2010

SENSITIVITY ANALYSIS – THE EFFECTS OF GLASGOW OUTCOME SCALE MISCLASSIFICATION ON TRAUMATIC BRAIN INJURY CLINICAL TRIALS

Juan Lu
Virginia Commonwealth University

SENSITIVITY ANALYSIS

– THE EFFECTS OF GLASGOW OUTCOME SCALE MISCLASSIFICATION ON TRAUMATIC BRAIN INJURY CLINICAL TRIALS

Three manuscripts submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

by

Juan Lu
M.D., School of Medicine of Suzhou University, China, 1983
M.P.H., Virginia Commonwealth University, 2001
Ph.D., Virginia Commonwealth University, 2010

Director: Kate L. Lapane Ph.D.
Professor and Chair, Department of Epidemiology and Community Health

Virginia Commonwealth University
Richmond, Virginia
April, 2010

Acknowledgment

I would like to dedicate the research project to the late Dr. Anthony Marmarou who was my mentor and supervisor for the last ten years, for his insight into the project, as well as for his support throughout the program. I would like to thank my advisor, Dr. Lapane, for her training and guidance, as well as my committee members, Dr. Turf, Dr. Ward, and Dr. Cifu for their help. I would also like to thank my family, for their support and patience during the years it has taken me to graduate. Last, but not least, I would like to thank my colleagues from the International Mission for Prognosis And Clinical Trial (IMPACT) project for allowing me to co-author with them.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

A METHOD FOR REDUCING MISCLASSIFICATION IN THE EXTENDED GLASGOW OUTCOME SCORE

IMPACT OF MISCLASSIFICATION ON THE ORDINAL GLASGOW OUTCOME SCALE IN TRAUMATIC BRIAN INJURY CLINICAL TRIALS

A METHOD FOR REDUCING MISCLASSIFICATION IN THE EXTENDED GLASGOW OUTCOME SCORE

ABSTRACTS

SENSITIVITY ANALYSIS

– THE EFFECTS OF GLASGOW OUTCOME SCALE MISCLASSIFICATION ON TRAUMATIC BRAIN INJURY CLINICAL TRIALS

By Juan Lu, M.D., M.P.H., Ph.D.

Three manuscripts submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at Virginia Commonwealth University.

Virginia Commonwealth University, 2010

Director: Kate L. Lapane Ph.D.
Professor and Chair, Department of Epidemiology and Community Health

I. EFFECTS OF GLASGOW OUTCOME SCALE MISCLASSIFICATION ON TRAUMATIC BRAIN INJURY CLINICAL TRIALS

The Glasgow Outcome Scale (GOS) is the primary endpoint for efficacy analysis of clinical trials in traumatic brain injury (TBI). Accurate and consistent assessment of outcome after TBI is essential to the evaluation of treatment results, particularly in the context of multicenter studies and trials. The inconsistent measurement or interobserver variation on GOS outcome, or for that matter, on any outcome scales, may adversely affect the sensitivity to detect treatment effects in clinical trial. The objective of this study is to examine effects of nondifferential misclassification of the widely used five-category GOS outcome scale and in particular to assess the impact of this misclassification on detecting a treatment effect and statistical power. We followed two approaches. First, outcome differences were analyzed before and after correction for misclassification using a dataset of 860 patients with severe brain injury randomly sampled from two TBI trials with known differences in outcome. Second, the effects of misclassification on outcome distribution and statistical power were analyzed in simulation studies on a hypothetical 800-patient dataset. Three potential patterns of nondifferential misclassification (random, upward and downward) on the dichotomous GOS outcome were analyzed, and the power of finding treatments differences was investigated in detail. All three patterns of misclassification reduce the power of detecting the true treatment effect and therefore lead to a reduced estimation of the true efficacy. The magnitude of such influence not only depends on the size of the misclassification, but also on the magnitude of the treatment effect. In conclusion, nondifferential misclassification directly reduces the power of finding the true treatment effect. An awareness of this procedural error and methods to reduce misclassification should be incorporated in TBI clinical trials.

II. IMPACT OF MISCLASSIFICATION ON THE ORDINAL GLASGOW OUTCOME SCALE IN TRAUMATIC BRIAN INJURY CLINICAL TRIALS

The methods of ordinal GOS analysis are recommended to increase efficiency and optimize future TBI trials. To further explore the utility of the ordinal GOS in TBI trials, this study extends our previous investigation regarding the effect of misclassification on the dichotomous GOS to examine the impact of misclassification on the 5-point ordinal scales. The impact of nondifferential misclassification on the ordinal GOS was explored via probabilistic sensitivity analyses using TBI patient datasets contained in the IMPACT database (N=9,205). Three patterns of misclassification including random, upward and downward patterns were extrapolated, with the pre-specified outcome classification error distributions. The conventional 95% confidence intervals and the simulation intervals, which account for the misclassification only and the misclassification and random errors together, were reported. Our simulation results showed that given a specification of a minimum of 80%, modes of 85% and 95% and a maximum of 100% for both sensitivity and specificity (random pattern), or given the same trapezoidal distributed sensitivity but a perfect specificity (upward pattern), the misclassification would have caused an underestimated ordinal GOS in the observed data. In another scenario, given the same trapezoidal distributed specificity but a perfect sensitivity (downward pattern), the misclassification would have resulted in an inflated GOS estimation. Thus, the probabilistic sensitivity analysis suggests that the effect of nondifferential misclassification on the ordinal GOS is likely to be small, compared with the impact on the binary GOS situation. The results indicate that the ordinal GOS analysis may not only gain the efficiency from the nature of the ordinal outcome, but also from the relative

smaller impact of the potential misclassification, compared with the conventional binary GOS analysis. Nevertheless, the outcome assessment following TBI is a complex problem. The assessment quality could be influenced by many factors. All possible aspects must be considered to ensure the consistency and reliability of the assessment and optimize the success of the trial.

III. A METHOD FOR REDUCING MISCLASSIFICATION IN THE EXTENDED GLASGOW OUTCOME SCORE

The eight-point extended Glasgow Outcome Scale (GOSE) is commonly used as the primary outcome measure in traumatic brain injury (TBI) clinical trials. The outcome is conventionally collected through a structured interview with the patient alone or together with a caretaker. Despite the fact that using the structured interview questionnaires helps reach agreement in GOSE assessment between raters, significant variation remains among different raters. We introduce an alternate GOSE rating system as an aid in determining GOSE scores, with the objective of reducing inter-rater variation in the primary outcome assessment in TBI trials. Forty-five trauma centers were randomly assigned to three groups to assess GOSE scores on sample cases, using the alternative GOSE rating system coupled with central quality control (Group 1), the alternative system alone (Group 2), or conventional structured interviews (Group 3). The inter-rater variation between an expert and untrained raters was assessed for each group and reported through raw agreement and with weighted kappa (k) statistics. Groups 2 and 3 without central review yielded inter-rater agreements of 83% (weighted k¼0.81; 95% CI 0.69, 0.92) and 83% (weighted k¼0.76, 95% CI 0.63, 0.89), respectively, in GOS scores. In GOSE, the groups had an agreement of 76% (weighted k¼0.79; 95% CI 0.69, 0.89), and 63% (weighted k¼0.70; 95% CI 0.60, 0.81), respectively. The group using the alternative rating system coupled with central monitoring yielded the highest inter-rater agreement among the three groups in rating GOS (97%; weighted k¼0.95; 95% CI 0.89, 1.00), and GOSE (97%; weighted k¼0.97; 95% CI 0.91, 1.00). The alternate system is an improved GOSE rating method that reduces inter-rater variations and provides for the first time, source documentation and structured narratives that allow a thorough central

review of information. The data suggest that a collective effort can be made to minimize inter-rater variation.

EFFECTS OF GLASGOW OUTCOME SCALE MISCLASSIFICATION ON TRAUMATIC BRAIN INJURY CLINICAL TRIALS

Introduction

Accurate and consistent assessment of outcome after TBI is essential to the evaluation of treatment results, particularly in the context of multicenter studies and trials. Various studies have investigated inter-observer agreement and misclassification of TBI outcome measures commonly used in TBI studies, and in general found that interobserver variation or misclassification on GOS outcome does exist (Anderson et al., 1993; Brooks et al., 1986; Choi et al., 2002; Maas et al., 1983; Marmarou, 2001; Pettigrew et al., 2003; Scheibel et al., 1998; Teasdale et al., 1998; Wilson et al., 1998; Wilson et al., 2002; Wilson et al., 2007), ranging from 17% (Marmarou, 2001) to 40% (Wilson et al., 2007) in practices. Previous work has shown that this could attenuate the true treatment effect and reduces the power of detecting the efficacy of treatment (Choi et al., 2002). However, little is known on how different misclassification directions or patterns might affect analysis of treatment effects in double-blinded TBI trials. It would seem reasonable to suspect that misclassification in a clinical trial would possibly effect both the treatment and control groups equally. However, even in this case, there is a profound effect on the analysis of the treatment effects (Choi et al., 2002). In clinical practice, nondifferential misclassification may affect the GOS outcome through three potential patterns: The *random* pattern refers to the misclassification between the adjacent categories that have an equal rate or chance of being classified for both treatment groups. The *upward* pattern means more true outcome categories are classified into better outcome categories for both groups; the *downward* pattern means more true outcome categories are classified into less optimistic outcome categories for both groups. The objective of this study is to

investigate whether these three potential patterns of measurement error may have differential effects on the power of finding treatment differences in double blinded TBI trials.

Materials and Methods

Misclassification

Misclassification in this paper is defined as an incorrect classification of the GOS outcome in TBI trials. Furthermore, for the purpose of discussing the outcome analysis of a double-blinded TBI trial, we assume in this study that the rates of misclassification are the same for both treatment and control groups. Thus, the outcome misclassification discussed in this study is *nondifferential or random*, and as defined above includes three potential patterns: (1) random, (2) upward and (3) downward for both treated and control groups. Realizing that misclassification may be a combination of upward and downward grading in either the placebo or treatment group, we selected patterns, which combined both directions of misclassifications. More specifically, we defined the "upward" pattern as 20 % of patients in both control and treated groups misclassified to a higher outcome category and 10 % of patients misclassified in a lower outcome category. The downward pattern was defined as 20 % of patients misclassified in a lower outcome category while 10 % of patients were misclassified to a higher category. These hypothetical percentages of misclassification are in the range of GOS misclassification found in other studies (Anderson et al., 1993; Maas et al., 1983; Marmarou, 2001; Wilson et al., 1998). Our focus in this report is to study misclassification applied equally to placebo and treated groups. However, an imbalance or non-random misclassification among treated and control groups is not considered in this report. Among all five categories of GOS outcome [Death (D), Vegetative (V), Severe Disability (SD), Moderate Disability (MD) and Good Recovery (GR)], only the category of death can be excluded from

misclassification, whereas the other four categories are all subject to misclassification, albeit to a different degree. To study the effect of misclassification, it is assumed that a certain rate of misclassification exists in a patient's outcome in two adjacent categories.

Patient Data

For analysis of the effect of misclassification on analysis of outcome differences, we used a dataset of 860 patients with severe brain injury randomly sampled from 2 TBI trials with known differences in outcome (Hukkelhoven et al., 2002).

For a more detailed analysis of the effect of misclassification on outcome distribution and statistical power we used a hypothetical 800 patients dataset (400 patients in each arm). In this dataset a 55% favorable outcome and a 20% mortality outcome distribution was considered as baseline. For both approaches the GOS was dichotomized into favorable (GR/MD) versus unfavorable (SD/V/D).

Statistical Method

Three patterns of misclassification on dichotomized GOS were studied: 1) random pattern: 20% GOS outcomes being equally misclassified between favorable and unfavorable outcome categories for both study groups, 2) upward pattern: 20% unfavorable outcomes being misclassified into favorable, and 10% favorable into unfavorable for both study groups, and 3) downward pattern: 20% favorable outcomes being misclassified into unfavorable, and 10% unfavorable into favorable for both study groups. For a dichotomous GOS outcome (GR/MD vs. SD/V/D), the simulated misclassification rates were only applied among survivors (i.e., between GR/MD and SD/V), however, all outcomes, including death, were assessed in the final outcome distribution measurement [i.e., (number of favorable outcomes/treatment total) - (number of favorable outcomes /control total).

Power Calculation

In this study, the power was defined as the probability of finding the difference between the treatment and control groups with a 95% two-sided significance. The calculation was based on a range of hypothetical two-proportion comparisons. No covariates were considered to simplify the problem. The powers, under a hypothetical condition with no misclassification and three simulated cases with misclassification, were compared.

The treatment effect in the hypothetical dataset was created following a conventional method (Bolland et al., 1998). For example, 10% treatment effect on a dichotomous GOS outcome [favorable (GR/MD) vs. unfavorable (SD/V/D)] was defined as an overall 10% outcome shift from the unfavorable to the favorable outcome in the treatment group, i.e. the favorable outcome in the treatment group increased by 10% and the unfavorable outcome decreased by 10% from the baseline.

Further, recognizing no misclassification on the outcome of death within the unfavorable outcome category, we applied the hypothetical treatment effect into the outcome of death, and the remaining unfavorable outcomes (i.e. a combined SD and V) individually. For the outcome of death, 10% treatment effect was defined as a 10% absolute reduction of the baseline numbers, and it was assumed that 10% of patients' outcomes were improved from death to better outcome categories including V and SD. Finally, the remaining numbers of unfavorable outcomes (SD/V) equaled the total treatment group minus 10% of the increased baseline favorable outcome numbers and minus 10% deducted baseline death numbers. The two-sided Chi-Square test was used for the dichotomous outcome comparisons.

Results

Effect of Misclassification

The effects of misclassification on the dichotomous outcome estimation were demonstrated by an actual phase III TBI trial patient dataset displayed in table 1. It was assumed that there were certain rates of outcome categories being misclassified. Thus, reversing the hypothetical misclassified outcome numbers to the observed outcome data would be helpful in gauging the effect of misclassification on the outcome analysis and the three possible misclassification models were applied.

Table 1. Effect of Misclassifications on the Observed Dichotomous GOS Outcomes

		Observed dichotomous GOS[1]			GOS after misclassification corrections[2]								
					Random 20% up and 20% down			*Upward* 20% up and 10% down			*Downward* 10% up and 20% down		
		Unfav.		Fav.	Unfav.		Fav.	Unfav.		Fav.	Unfav.		Fav.
Groups	N	D	V/SD	MD/G	D	V/SD	MD/G	D	V/SD	MD/G	D	V/SD	MD/G
Treatment	430	93	85	252	93	29	308	93	73	264	93	25	312
Control	430	131	81	218	131	35	264	131	73	226	131	30	269
Difference (%)[3]		7.9			10.2			8.8			10.0		
P-value[4]		0.020			0.002			0.009			0.002		

1. Observed Glasgow Outcome Scale: D=Death, V=Vegetative, SD=Severe Disabled, MD=Moderate Disabled, G=Good Recovery

2. The corrected GOS misclassifications are given by the equation: Fav(Observed) = Fav(True) - Rate1*Fav(True) + Rate2*[N - D - Fav(True)]. Where 1) Fav(Observed) is the count of the observed favorable outcomes, 2) Fav(True) is the count of the corrected favorable outcomes, 3) Rate 1 and Rate 2 are the rates of upward and downward misclassification respectively, and 4) N and D represent the group total and the number of death respectively. For example, after correcting for 20% upward and 20% downward misclassification, the equation 252 = X - 0.2*X + 0.2*(430 - 93 - X) gives the corrected MD/G=308, and V/SD=430-93-308=29 for the treatment group; while equation 218 = X - 0.2*X + 0.2*(430 - 131 - X) gives corrected MD/G=264 and V/SD=430-131-264=35 for the control group.

3. Difference (%) in the favorable outcomes between the treatment and control groups.

4. Chi-Square Test (two-sided)

Random Pattern

In the random pattern, the adjacent outcome categories have an equal rate of being misclassified for both treatment and control groups. For example, in Table 1, it was assumed that equal rates (20%) of patients had been misclassified as favorable or unfavorable outcome for both groups. If these misclassified outcome numbers were corrected based on our assumptions, the true underlying number of patients with the favorable outcomes would be 308 for the treatment group,

264 for the control, and the percentage difference in favorable outcomes between the two groups would be (308 –264) / 430 or 10.2 % (p-value=0.002). The method for calculation is shown in the Table 1 legend. Before the 20% misclassification correction, the observed difference is 7.9 and p-value is 0.02. Thus, misclassification introduces an error of 2.3 % (10.2-7.9).

Upward Pattern

The upward model resulted in an upward trend of misclassification for both treatment and control groups, where the rate of patients being misclassified was higher (20 %) from the unfavorable outcomes to the favorable outcomes than the rate exchange from the other direction (10%). If the misclassified outcome numbers were corrected, the number of patients with the favorable outcomes would be 264 for the treatment group, and 226 for the control. The actual percentage difference in favorable outcomes between the two groups would be 8.8 (p-value=0.009) instead of the observed difference of 7.9 (p-value=0.02). In this case, misclassification introduces an error of 0.9 % (8.8 – 7.9).

Downward Pattern

In the downward model, the rate of being misclassified was lower (10%) from the unfavorable outcomes to the favorable outcomes than the rate exchange from the other direction (20%) for both treatment and control groups. After the misclassified outcome numbers were corrected, the numbers of patients with the favorable outcome would be 312 for the treatment group, and 269 for the control. The percentage difference in favorable outcomes between the two groups would be 10.0 (p-value=0.002) resulting in misclassification error of 2.1 % (10.0 – 7.9).

Thus, corrections for all three patterns of misclassification demonstrated a potential for *greater* outcome differences and *smaller* p-values than the observed dataset if the study assumption was true and the misclassification existed in the observed outcome measurement.

Misclassification and Outcome Distribution

Table 2 illustrates the relationship between misclassification and the dichotomous outcome distribution under three misclassification models. A hypothetical 800-patient dataset (400 patients each group) with a 55% favorable outcome rate and 20% mortality rate was used for this illustration.

Table 2. Effect of Misclassification on the Binary GOS Outcome Distribution

	Simulated treatment effect (%)	0% Misclassification				GOS outcome misclassifications: *Random* 20% up and 20% down					*Upward* 20% up and 10% down					*Downward* 10% up and 20% down				
		D	V/SD	MD/G	Outcome %Dif.[1]	D	V/SD	MD/G	Outcome %Dif.	%Dec.[2]	D	V/SD	MD/G	Outcome %Dif.	%Dec.	D	V/SD	MD/G	Outcome %Dif.	%Dec.
Control		80	100	220		80	124	196			80	102	218			80	134	186		
Treatment	***0%***	80	100	220	0	80	124	196	0		80	102	218	0		80	134	186	0	
	5%	76	84	240	5	76	115	209	3.2	1.8	76	91	233	3.7	1.3	76	124	200	3.6	1.4
	10%	72	68	260	10	72	106	222	6.4	3.6	72	80	248	7.4	2.6	72	113	215	7.2	2.8
	15%	68	52	280	15	68	98	234	9.6	5.4	68	70	262	11.1	3.9	68	103	229	10.8	4.2

1. Outcome difference (%) 2. Deduction (%) from the expected outcome difference

In general, without an outcome difference (i.e., 0% treatment effect), the misclassified outcome numbers were the same for both treatment and control groups and the misclassification only resulted in outcome distribution shifts, but not in outcome differences for all three models. However, with an outcome difference (i.e., 5%, 10%, 15% treatment effect), the outcome distributions for the treatment group were different from the distributions for the control group. As a result, the misclassified outcome numbers for the treatment and control groups were also different. For example, a 20% outcome number exchange between the favorable and unfavorable outcome categories in the random misclassification case caused the 5%, 10% and 15% outcome differences to decrease to 3.2%, 6.4% and 9.6%. The reduction is 1.8%, 3.6% and 5.4% respectively from the previous outcome differences.

Similarly, in the *upward* (20% up and 10% down) and *downward* (20% down and 10% up) misclassification examples, after applying the rate exchange between the dichotomous outcome categories, the outcome differences (i.e., 5%, 10%, 15%) decreased to 3.7%, 7.4% and 11.1% (Upward), as well as 3.6%, 7.2% and 10.8% (Downward), which were 1.3%, 2.6% and 3.9% (Upward), as well as 1.4%, 2.8%, and 4.2% (Downward) reductions from the original outcome differences respectively.

Thus, it is conceivable that the impact from a misclassification on a dichotomous outcome measurement is not only related to the misclassification but also depends on the outcome distributions of the two study groups. Regardless of the random, upward and downward patterns of misclassifications, for a fixed rate of misclassification, the dichotomous outcome difference depends on the size of treatment effect or the difference in outcome distribution between the treatment and control groups. This is illustrated in Table 2, where all three misclassification examples have revealed that the more the treatment group differs from the control, the greater the impact of misclassification.

Misclassification and Power

The powers of detecting the expected treatment effect and the misclassified treatment effect were compared and are illustrated in Figure 1a to 1c using the same hypothetical 800 patient dataset.

Figure 1a. Effect of Random Misclassification on the Power

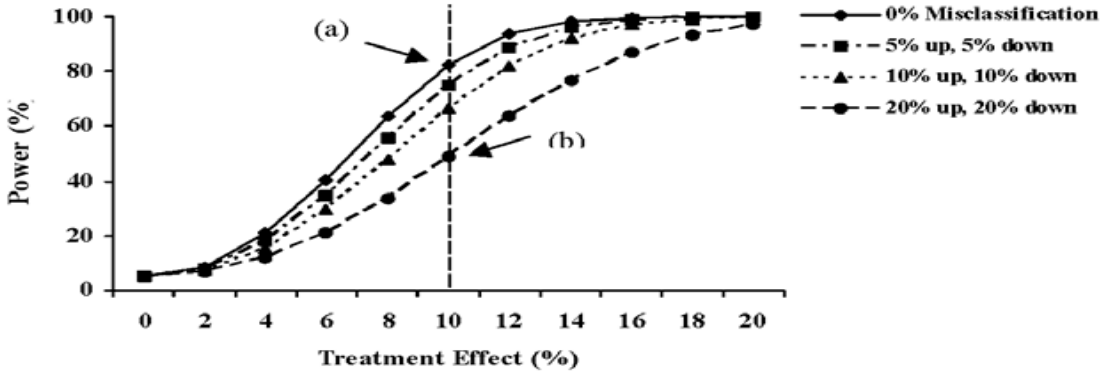


Figure 1a shows the effect of random misclassification pattern on the power. The solid line presents the correlation between the power and the expected treatment effect and dashed lines present the correlation between the random misclassified treatment effect and power according to the symbol key. For example, for the case of a 10 % treatment effect, a 20 % up and down random misclassification would result in a reduction of power from 82 % (point a) to 49 % (point b) thereby rendering the trial non-significant.

Figure 1a. Effect of Random Misclassification on the Power

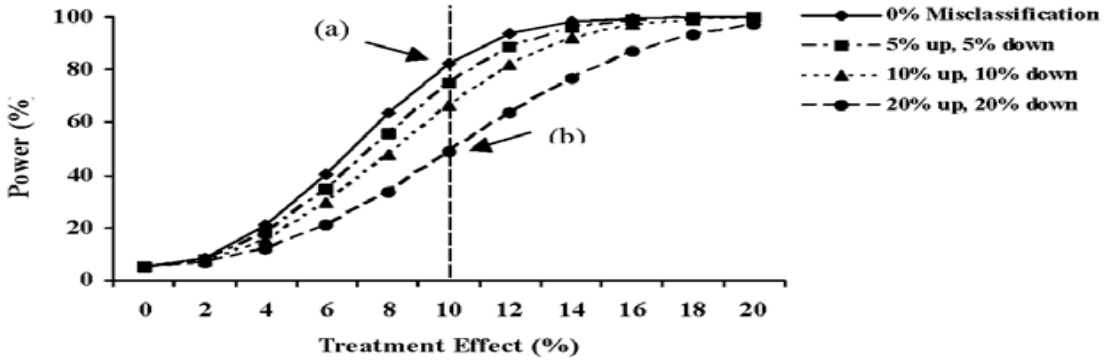


Figure 1b shows the effect of upward misclassification pattern on the power. The solid line represents the correlation between the power and the expected treatment effect and dashed lines represent the correlation between the upward misclassified treatment effect and power according to the symbol key. For example, for the case of a 10 % treatment effect, a 20 % up and 10 % down (lowest dashed line) misclassification results in a reduction of power from 82 % (point a) to 60 % (point b).

Figure 1c. Effect of Downward Misclassification on the Power

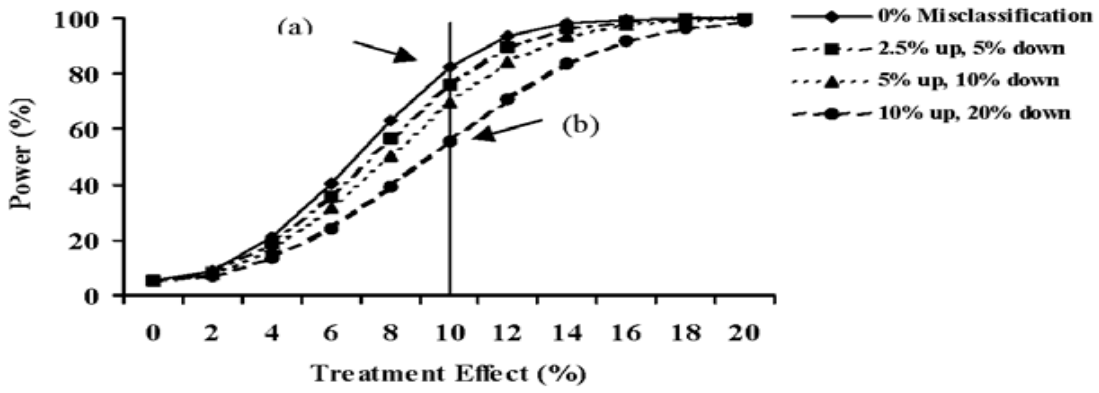


Figure 1c shows the effect of downward misclassification pattern on the power. The solid line represents the correlation between the power and the expected treatment effect and dashed lines represent the correlation between the downward misclassified treatment effect and power according to the symbol key. For example, for the case of a 10 % treatment effect, a 10 % up and 20 % down (lowest dashed line) misclassification results in a reduction of power from 82 % (point a) to 55 % (point b).

An example of the effect of random pattern on the power is shown in Fig 1a. Under a given treatment effect (i.e., the improved proportion of the favorable outcome in the treatment group), the power was inversely associated with the rate of the misclassification. For example, the power of detecting a 10% true treatment effect with a two-sided 95 percent significance is 82% (solid line), however, after the 5%, 10% and 20% misclassifications were applied to the expected treatment effect, the power of detecting the same 10% treatment difference decreased to 75%, 67% and 49% respectively. Clearly, this was due to the altered outcome difference by the misclassification. *The higher the misclassification rate, the smaller the treatment effect and the lower the power.*

Figure 1b and 1c demonstrate the effect of the upward and downward misclassification patterns on the power. Similar results on reducing the power were observed, albeit in different

degrees. If using a power to detect a 10% treatment effect with a two-sided 95% significance as an example, the upward pattern with a combination of 5% up and 2.5% down, 10% up and 5% down and 20% up and 10% down were considered to be the rate exchange between the dichotomous outcome categories, then the desired 82% power would be decreased to 78%, 72% and 60% accordingly. On the other hand, if the situation was reversed, namely downward pattern with 2.5% up and 5% down, 5% up and 10% down and 10% up and 20% down were used as the rate exchanges between the outcome categories, then the expected 82% power would be reduced to 76%, 70% and 55%, respectively.

Discussion

Outcome measurements and outcome misclassification in trials of head injury

The GOS is widely used for TBI outcome measurement (Jennett and Bond, 1975), and recommended as primary endpoint for assessing efficacy of novel therapeutic approaches in clinical trials. For purposes of analysis in clinical trials, the GOS is commonly dichotomized into favorable versus unfavorable outcome, collapsing the 5-point categorical outcome scale into a binary outcome measure (Bullock et al., 2002; Choi et al., 1998; Maas et al., 1997; Narayan et al., 2002; Teasdale et al., 1998; Wilson et al., 2002). Despite the acceptance of the GOS as a global functional outcome measure, it has been criticized as being insensitive, especially in the more favorable end of outcome (Bullock et al., 2002; Levin et al., 2001; Teasdale et al., 1998). The 8-point extended GOS (GOSE) has been introduced to increase sensitivity of outcome assessment, and the use of a structured interview is advocated to obtain more consistency in outcome assignment (Fayol et al., 2004; Wilson et al., 1998). Although the GOSE offers increased sensitivity, this benefit may be offset by a

higher rate of misclassification. Recent evidence indicates an agreement rate as low as 60% in GOSE by untrained investigators (Wilson et al., 2007).

Misclassification, especially the nondifferential misclassification, is a relevant issue in clinical trial design. Previous work indicated that random misclassification could mask the true efficacy and reduce the power of finding a treatment effect (Choi et al., 2002). By understanding the consequence of outcome misclassification, efforts could be made to improve the accuracy and consistency of outcome measurements.

The present study has confirmed the substantial effects of nondifferential misclassification on outcome analysis and statistical power under various scenarios. *One may question whether the effects of misclassification are substantial enough to be important. Clearly, from our analysis, we have found that a treatment effect may be reduced from 10% to 6.8% by a 20% random misclassification* (i.e. 20% up and 20% down)*, which is more than sufficient to render a trial ineffective.* The effect of misclassification on treatment effect is summarized in Table 3.

Table 3. Reduction of Treatment Effect and Power by Misclassification

Patterns of Misclassification[1]	Treatment effect Reduction	Power Reduction[2]
Random		
10% up/down	10% → 8.4%	80% → 66.5%
20% up/down	10% → 6.8%	80% → 48.6%
Upward		
10% up	10% → 9.2%	80% → 76.6%
20% up	10% → 8.4%	80% → 69.9%

1. Misclassification on both treatment and control arms, assume 55% favorable outcome and 20% mortality
Two Arm Trial, N=800, Expected Treatment Effect=10%, Power=80% and 95% two-sided significance

Moreover, the scenarios and the rates of misclassification investigated are not unrealistic to clinical practice. Marmarou (Marmarou, 2001) conducted a study within the American Brain Injury Consortium to ascertain the reliability of the GOS rating and found an upward shift of 17.4 percent of severe patients to the moderate disability category. An upward shift of outcome assignment had been previously reported (Anderson et al., 1993) and is a likely result of the

optimism of a patient's primary care providers who compare the improved outcome to the serious condition immediately after injury, rather than to the healthy pre-injury status. Conversely, a rigid application of the criteria from the structured interview or questionnaires by research workers tends to allocate patients to lower outcome categories (Teasdale et al., 1998; Wilson et al., 2007). Therefore, nondifferential misclassification may be found in either the upward or downward direction, based on different clinical scenarios. It is for this reason, that three patterns were studied in this report.

An Assessment of three possible misclassification patterns

To demonstrate the effect of nondifferential misclassification on the outcome, we applied three possible patterns of misclassification on a real Phase III head injury trial data using 5-category GOS outcome distribution in Table 1. According to the results from previous studies, we assume that there were certain rates of nondifferential misclassification embedded in the observed dataset, we corrected the hypothetical misclassified outcome numbers to the observed data using three models. After the numbers were corrected, a larger outcome difference and a smaller p-value were revealed in all three misclassification patterns.

Therefore, if the misclassification indeed existed in the past trial dataset as described in this study and as suggested by other studies, the true outcome difference would have been larger. More importantly, our study indicated that regardless of which direction dichotomous outcome was misclassified (i.e., random, upward and downward), *the effect of nondifferential misclassification always tends to reduce the true dichotomous outcome difference.*

It should be noted that the random and the downward patterns in our examples seemed to have a larger effect on reducing the outcome difference than the upward pattern. This is likely due to the outcome distribution being misclassified and the rates of misclassification being applied. For

example, more outcome numbers were exchanged from the category of MD/GR (i.e., 20% MD/GR = (0.2)*(252) = 50) with the numbers of V/SD in the random or downward cases, as compared with the numbers that were exchanged (10% MD/GR = (0.1)*(252) = 25) in the upward case. Thus, it is reasonable to understand why the random and the downward patterns had a larger impact on the outcome difference in our example.

In summary, the true outcome difference is always affected more by a higher misclassification rate and a larger difference in outcome distributions between the treatment and control groups. Therefore, any procedures that minimize the misclassification, such as proper outcome measurement techniques and the methods for improving the inter-observer agreement, should be indicated according to this study. Experience in the recent Phase III clinical trial on Dexanabinol showed that training of outcome assessors can be highly effective (Wilson et al., 2007).

Differential effects of Misclassification in Treatment and Control Groups

Although it is generally assumed that the rate of the misclassification under a blinded clinical trial condition is the same for both control and treatment groups (i.e., nondifferential or random misclassification), the effect of the misclassification on these two are unlikely to be the same in the presence of a treatment effect. This is a consequence of the different outcome distribution between treatment groups, as illustrated in Table 2.

For example, in the random misclassification (20% up/down) case, with no treatment effect, the misclassified outcome numbers are the same for both treatment and control groups; the misclassification only resulted in an outcome distribution shift but not in an outcome difference. However, with a treatment effect (i.e., 5%, 10%, 15%), the misclassified outcome numbers for the treatment and control groups are no longer the same, i.e., more patients' outcome in the treatment

group are affected by the misclassification due to a larger outcome difference. Thus, instead of having an 5%, 10%, 15% in treatment effect, only a 3.4%, 6.8%, 10.2% outcome difference results, which represents a 1.6%, 3.2% and 4.8% reduction of the previous outcome difference respectively.

The other two patterns followed a similar trend as well. Figure 2 shows the example of correlation between upward misclassifications and reduction of treatment effect using same hypothetical 800 patient data as in Table 2.

Figure 2. Reduction of Treatment Effect by Upward Misclassification

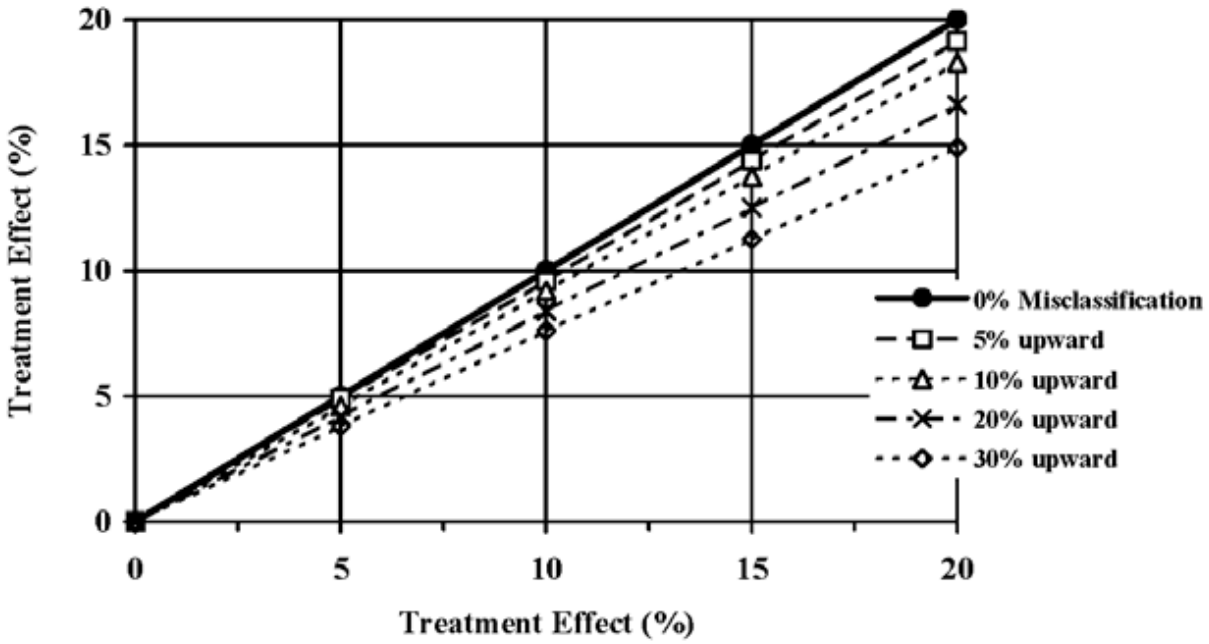


Figure 2 illustrates the reduction of treatment effect by upward outcome misclassification. The solid line represents the expected treatment effect and dashed lines represent the reduction of treatment effect by the upward misclassifications according to the symbol key. For example, for the case of a 10 % treatment effect, a 30 % up (lowest dashed line) misclassification results in a reduction of treatment effect to 7.6%, which is 2.4% reduction from the expected 10% treatment effect.

For a fixed rate of misclassification, the outcome difference depends on the size of treatment effect or the difference in outcome distribution between the treatment and control groups. For example, after a 20% upward misclassification, the expected 5%, 10%, 15% and 20% outcome

differences were reduced to 4.2%, 8.4%, 12.2% and 16.6% respectively. On the other hand, for a fixed treatment effect, the effect of misclassification on outcome difference depends on the rate of misclassification. For example, after 5%, 10%, 20% and 30% biased upward misclassification, an expected 10% treatment effect was reduced to 9.4%, 9.2%, 8.4%, and 7.6% respectively.

The implication here is that if a study drug does have an effect on improving the patients' outcome, the treatment group is likely to be affected more by the misclassification than the control group. The more a treatment differs from the control, the greater the number of patients affected by the misclassification, leading to a greater reduction in the true outcome difference.

It is important to note that the demonstration on the dichotomous outcome distribution can also be applied to more than two category distributions. For example, if there is a larger difference between MD and GR, the true difference between these two categories will be affected more by misclassification. Likewise, if a larger difference exists between SD and MD, the actual difference between these two will be decreased more as a result of the misclassification. This topic will be studied in greater detail in the future.

Dealing with misclassification

Since all GOS categories can be misclassified except death and as the affected outcome numbers are associated with the treatment effect and/or the outcome distribution, one might relate the issue to the choice of outcome measurements in head injury clinical trials. One study has suggested that an increase in outcome categories leads to an increase in misclassifications, and several other studies have proved that inter-observer disagreement is much higher in the 8-category GOS than that in the 5-category GOS (Choi et al., 2002; Maas et al., 1983). These observations underline the notion that the outcome measurement with fewer outcome categories might be less

affected by the misclassification. A careful balance will need to be sought between the desire for more sensitive expanded outcome measures and adverse effects of misclassification.

We suggest that both outcome misclassification and the sensible outcome measurement are important issues in the TBI trial design, which, in turn, is directly associated with the success of a trial. However, both the strategy to minimize the outcome misclassification, and to select a sensible outcome measurement should be considered separately. Although outcome misclassification is unavoidable, it is possible that errors in classification may be reduced. Accordingly, procedures such as structured interviews, proper outcome information resources, quality assurance of outcome evaluation and properly trained personnel have been previously shown to be successful approaches of minimizing the misclassification (Marmarou, 2001; Pettigrew et al., 2003; Wilson et al., 1998; Wilson et al., 2007). These measures as well as developing new strategies are recommended in the clinical trial design.

On the other hand, carefully examining the outcome distribution from Phase II trials and selecting a sensitive outcome measurement to match each individual outcome distribution should be considered. For instance, if a treatment effect was mainly focused between the moderate and severe disability categories or other adjacent GOS categories, a dichotomous outcome would be a better choice over more GOS categories (Choi et al., 2002). However, if more or all categories of the GOS are affected by the treatment, then the dichotomous GOS would be less powerful than using more GOS categories

Effect of Misclassification on power and sample size

Recognizing that outcome misclassification has a significant potential to reduce the true treatment effect, one would naturally relate this consequence to the power and sample size of a trial design. For a typical Phase III TBI trial, a sample size of 800 patients (i.e. 400 patients in treatment

group, 400 in placebo group) is required in order to detect an absolute treatment effect that increases the proportion of favorable outcomes from 50% to 60% with 80% power and 5% significance. We used a similar design to examine the effect of outcome misclassification on the desired power in the TBI trial. The correlation between the power and three potential patterns of misclassification was depicted in Figures 1a through 1c.

As one might expect, in parallel with the effect of reducing the true outcome difference, all three patterns of misclassification have an inverse effect on the power. For instance (Figure 1a), without misclassification, the expected power of detecting a 10% treatment effect (i.e., improving favorable outcome from 55% to 65% in the treatment group in our example) was 82%; with the same condition and a 10% random (i.e. 10% up and 10% down) misclassification for both study groups, the power of detecting such effect decreased to 67%; similarly, the powers under the upward (i.e.10% up and 5% down), and the downward (i.e. 10% down and 5% up) misclassification condition reduced the power from 82 % to 72% and 70% respectively. Clearly, the examples shown in this study demonstrate that the desired power to detect the treatment effect could be compromised by misclassification of the dichotomous GOS outcome; the greater the number of outcomes misclassified, the greater the degree of power compromised.

Compensation for Reduced Power due to Misclassification

As misclassification reduces power, it would seem reasonable to simply increase the sample size to compensate for the power reduction. This can be done. However, increasing the sample size can only raise the power but cannot compensate for treatment effect due to misclassification. Using our previous example in Table 3., a 10% random misclassification can reduce the original 10% treatment effect to 8.4%, and the power was subsequently reduced from 80% to 66.5% for detecting 8.4% treatment effect. In this example, one can increase the sample size

from 800 to 1094 in order to raise the power from 66.5% to 80% for detecting 8.4% treatment effect, but still, the increased sample size cannot compensate the 1.6% (10%-8.4%) treatment reduction. This further emphasizes the importance of designing procedures to minimize the effect of misclassification. In summary, the only way to blunt the reduction of treatment effect is to reduce misclassification.

Conclusions

All three patterns of nondifferential misclassification act to attenuate the treatment effect and reduce the power of detecting the true treatment effect. In the case of a positive drug effect, misclassification leads to a conservative estimation of the true efficacy. The magnitude of such influence not only depends on the size of the misclassification, but also on the magnitude of treatment effect. Nondifferential misclassification directly reduces the power of finding the true treatment effect. If the outcome of the treatment arm is worse, then misclassification acts to blunt the difference between placebo and treatment. Thus, an awareness of this procedural error and methods to reduce misclassification should be incorporated in TBI clinical trials.

Financial Disclosure: None

Acknowledgement: Grant Support was provided by NS-042691 and NS019235-21

References

Anderson, S.I., Housley, A.M., Jones, P.A., Slattery, J., and Miller, J.D. (1993). Glasgow Outcome Scale: an inter-rater reliability study. Brain injury : [BI]. 7, 309-317.

Bolland, K., Sooriyarachchi, M.R., and Whitehead, J. (1998). Sample size review in a head injury trial with ordered categorical responses. Statistics in medicine. 17, 2835-2847.

Brooks, D.N., Hosie, J., Bond, M.R., Jennett, B., and Aughton, M. (1986). Cognitive sequelae of severe head injury in relation to the Glasgow Outcome Scale. Journal of neurology, neurosurgery, and psychiatry. 49, 549-553.

Bullock, M.R., Merchant, R.E., Choi, S.C., Gilman, C.B., Kreutzer, J.S., Marmarou, A., and Teasdale, G.M. (2002). Outcome measures for clinical trials in neurotrauma. Neurosurgical focus. 13, ECP1.

Choi, S.C., Clifton, G.L., Marmarou, A., and Miller, E.R. (2002). Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. Journal of neurotrauma. 19, 17-22.

Choi, S.C., Marmarou, A., Bullock, R., Nichols, J.S., Wei, X., and Pitts, L.H. (1998). Primary end points in phase III clinical trials of severe head trauma: DRS versus GOS. The American Brain Injury Consortium Study Group. Journal of neurotrauma. 15, 771-776.

Fayol, P., Carriere, H., Habonimana, D., Preux, P.M., and Dumond, J.J. (2004). French version of structured interviews for the Glasgow Outcome Scale: guidelines and first studies of validation. Annales de Readaptation et de Medecine Physique. 47, 142-156.

Hukkelhoven, C.W., Steyerberg, E.W., Farace, E., Habbema, J.D., Marshall, L.F., and Maas, A.I. (2002). Regional differences in patient characteristics, case management, and outcomes in traumatic brain injury: experience from the tirilazad trials. Journal of neurosurgery. 97, 549-557.

Jennett, B., and Bond, M. (1975). Assessment of outcome after severe brain damage. Lancet. 1, 480-484.

Levin, H.S., Boake, C., Song, J., Mccauley, S., Contant, C., Diaz-Marchan, P., Brundage, S., Goodman, H., and Kotrla, K.J. (2001). Validity and sensitivity to change of the extended Glasgow Outcome Scale in mild to moderate traumatic brain injury. Journal of neurotrauma. 18, 575-584.

Maas, A.I., Braakman, R., Schouten, H.J., Minderhoud, J.M., and van Zomeren, A.H. (1983). Agreement between physicians on assessment of outcome following severe head injury. Journal of neurosurgery. 58, 321-325.

Maas, A.I., Dearden, M., Teasdale, G.M., Braakman, R., Cohadon, F., Iannotti, F., Karimi, A., Lapierre, F., Murray, G., Ohman, J., Persson, L., Servadei, F., Stocchetti, N., and Unterberg, A. (1997). EBIC-guidelines for management of severe head injury in adults. European Brain Injury Consortium. Acta Neurochirurgica. 139, 286-294.

Marmarou, A. (2001). Head Trauma: Basic, Preclinical, Clinical Direction. First edn. Willey: New York, pps. 15

Narayan, R.K., Michel, M.E., Ansell, B., Baethmann, A., Biegon, A., Bracken, M.B., Bullock, M.R., Choi, S.C., Clifton, G.L., Contant, C.F., Coplin, W.M., Dietrich, W.D., Ghajar, J., Grady,

S.M., Grossman, R.G., Hall, E.D., Heetderks, W., Hovda, D.A., Jallo, J., Katz, R.L., Knoller, N., Kochanek, P.M., Maas, A.I., Majde, J., Marion, D.W., Marmarou, A., Marshall, L.F., McIntosh, T.K., Miller, E., Mohberg, N., Muizelaar, J.P., Pitts, L.H., Quinn, P., Riesenfeld, G., Robertson, C.S., Strauss, K.I., Teasdale, G., Temkin, N., Tuma, R., Wade, C., Walker, M.D., Weinrich, M., Whyte, J., Wilberger, J., Young, A.B., and Yurkewicz, L. (2002). Clinical trials in head injury. Journal of neurotrauma. 19, 503-557.

Pettigrew, L.E., Wilson, J.T., and Teasdale, G.M. (2003). Reliability of ratings on the Glasgow Outcome Scales from in-person and telephone structured interviews. The Journal of head trauma rehabilitation. 18, 252-258.

Scheibel, R.S., Levin, H.S., and Clifton, G.L. (1998). Completion rates and feasibility of outcome measures: experience in a multicenter clinical trial of systemic hypothermia for severe head injury. Journal of neurotrauma. 15, 685-692.

Teasdale, G.M., Pettigrew, L.E., Wilson, J.T., Murray, G., and Jennett, B. (1998). Analyzing outcome of treatment of severe head injury: a review and update on advancing the use of the Glasgow Outcome Scale. Journal of neurotrauma. 15, 587-597.

Wilson, J.T., Edwards, P., Fiddes, H., Stewart, E., and Teasdale, G.M. (2002). Reliability of postal questionnaires for the Glasgow Outcome Scale. Journal of neurotrauma. 19, 999-1005.

Wilson, J.T., Pettigrew, L.E., and Teasdale, G.M. (1998). Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. Journal of neurotrauma. 15, 573-585.

Wilson, J.T., Slieker, F.J., Legrand, V., Murray, G., Stocchetti, N., and Maas, A.I. (2007). Observer variation in the assessment of outcome in traumatic brain injury: experience from a multicenter, international randomized clinical trial. Neurosurgery. 61, 123-8; discussion 128-9.

IMPACT OF MISCLASSIFICATION ON THE ORDINAL GLASGOW OUTCOME SCALE IN TRAUMATIC BRIAN INJURY CLINICAL TRIALS

Introduction

Several recent studies (Bath et al., 2008; Optimizing Analysis of Stroke Trials (OAST) Collaboration 2007; McHugh et al., 2010) have explored the ordinal analysis of the 5-point ordinal Glasgow Outcome Scale (GOS) (Jennett and Bond, 1975) and other ordinal outcomes commonly used in Traumatic Brain Injury (TBI) and stroke clinical trials. Previously, studies have dichotomized GOS outcomes which split the GOS as an unfavorable (dead, vegetative status and severe disability) versus favorable (moderate disability and good recovery) outcome. Recent methodologic work indicates that analyzing ordinal outcomes as ordinal gives substantial gains over conventional dichotomized outcomes. Therefore, it is strongly recommended that both ongoing and future TBI and stroke trials using original ordinal outcomes and use methods of ordinal analyses.

To further explore the utility of the ordinal GOS in TBI trials, this study extends our previous investigation (Lu et al., 2008) regarding the effect of misclassification on the dichotomous GOS, to examine the impact of misclassification on the 5-point ordinal scales. In the previous study, we used a simple sensitivity analysis to explore three patterns of nondifferential misclassification on the dichotomous GOS and its impact on the efficacy analysis and statistical power. The results suggested that all three patterns of misclassification act to attenuate the treatment effect and reduce the statistical power. In the case of a positive drug effect, misclassification leads to a conservative estimation of true efficacy.

In this study, we explored the impact of nondifferential misclassification on the 5-point ordinal scales via a probabilistic sensitivity analyses (Fox et al., 2005; Lash and Fink, 2003). Probabilistic sensitivity analysis is an extension of simple sensitivity analysis in which the study assigns probability distributions to misclassification parameters, rather than a single value or a

series of discrete values within a range. By using a range of possible sensitivity and specificity parameters, the analysis calculates simulation intervals that portray the presumed uncertainty as more plausible misclassification. The final simulation intervals account for uncertainties from both outcome misclassification and random errors. The results will help investigators to better understand the impact of misclassification on the 5-point ordinal outcome in TBI trials in a quantitative manner.

Materials and Methods

Patient data

We used patient data from the IMPACT Database (Marmarou et al., 2007) as examples of typical selected head injury populations. The IMPACT project was an international collaboration linking researchers in Belgium, Netherlands, United Kingdom and United States of America, which was funded by National Institutes of Health (NIH) and aimed to develop methodologies to improve the design and analysis of clinical trials of head injury. The IMPACT database contains clinical data on 9,205 individual patients with moderate or severe head injury, from eight randomized controlled trials (RCTs) (n = 6,535) and three observational epidemiologic studies (n = 2,670). The patient data from RCTs represent the head injury population with more restricted inclusion criteria, whereas the data from the observational studies represent the population with more generalized characteristics.

For each individual study in the IMPACT Database, 400 subjects were randomly sampled with replacement, as the baseline samples or the placebo group. Another 400 subjects were randomly sampled from each placebo group with replacement, as the treatment group, respectively. Thus, each individual dataset was generated representing the general TBI population from RCTs and observational studies.

The treatment effect was simulated for the treatment group within each study based on the assumption that the effect of drug treatment followed a proportional odds model (McCullagh, 1980). The common odds ratio (COR) was calibrated so that there would be an overall 10% (COR=1.5) increase in the proportion of patients with a favorable outcome in the treatment group.

Nondifferential ordinal GOS outcome misclassification

The 5-point ordinal GOS includes categories of good (GR), moderate disability (MD), severe disability (SD), vegetative status (VS) and death (D). The categories of VS and D were combined and analyzed as one category in the study. In the context of double blind clinical trials, it was assumed that the misclassification on the outcome was non-differential, that is, the outcome misclassification was the same for both treated and control groups. Further, two assumptions were made to simulate the misclassification on the 5-point ordinal GOS outcome. First, the misclassification was made between two adjacent categories only, with a same set of biased parameters. For example, the misclassification was made between the categories of GR and MD and between MD and SD only, with a same sensitivity and specificity. Second, no misclassification was made for the category of VS, thus, for the combined category of VS and D.

Probabilistic sensitivity analysis on the ordinal outcome misclassification

In this study, we used the concept of the probabilistic sensitivity analysis introduced by Lash and Fink (2003) and Fox et al. (2005) and modified the SENSMAC (SAS Macros) by Fox and colleagues. The original SENSMAC was generated to provide probabilistic sensitivity analysis to quantify likely effects of misclassification of a dichotomous outcome, exposure or covariate. We modified the misclassification section of the SAS macros to explore the impact of ordinal GOS misclassification on TBI trials. The modification was mainly done for the category of MD where the misclassification could be made to both directions, that is, between MD and GR and between

MD and SD. Since the assumption was that the misclassification can only be made between two adjacent outcome categories, the reclassification for MD was performed through two independent binary situations as illustrated by the Figure 1. As a result, the terms of sensitivity and specificity were maintained for the two binary situations and the overall reclassified MD was recalculated accordingly.

Figure1. Illustration of the ordinal GOS outcome misclassification in which the category of MD could be misclassified to both adjacent categories

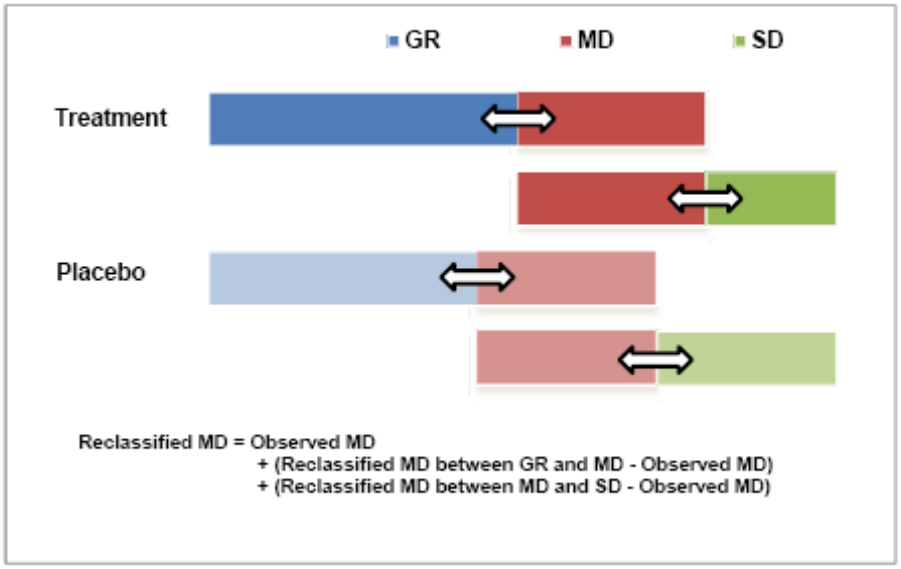


Figure 1 illustrates the reclassification process for the category of Moderate Disability (MD) for which the misclassification can be made to both directions. The study assumes the misclassification is between the two adjacent categories only, i.e., between the categories of Good Recovery (GR) and MD and of MD and Severe disability (SD). Thus, the reclassification for the category of MD is performed through two independent binary situations first, and then recalculated based on the observed MD, and the differences between the reclassified MD with GR and the observed MD and between the reclassified MD with SD and the observed MD.

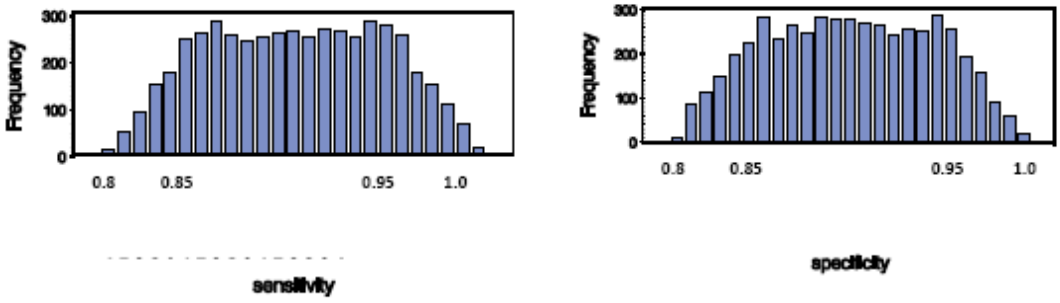
Three patterns of nondifferential misclassification

Three potential patterns of nondifferential misclassification were simulated to examine the impact of misclassifications on the 5-point ordinal GOS in TBI trials, including the patterns of random, upward and downward (Figure 2). For the random pattern, we specified a trapezoidal distribution (minimum of 80%, modes of 85 and 95%, and a maximum of 100%) for both

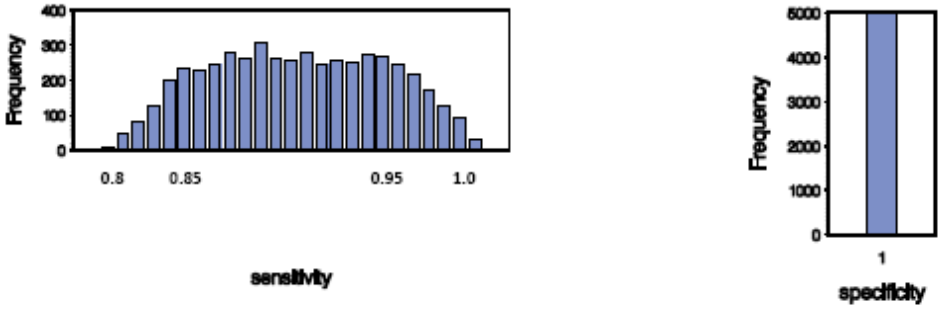
sensitivity and specificity; for the upward pattern, we specified trapezoidal distribution (minimum of 80%, modes of 85 and 95%, and a maximum of 100%) for sensitivity and a perfect specificity; while for the downward pattern, we specified a trapezoidal distribution (minimum of 80%, modes of 85 and 95%, and a maximum of 100%) for specificity and a perfect sensitivity.

Figure 2. Three patterns of misclassification

A. Random pattern



B. Upward pattern



C. Downward pattern

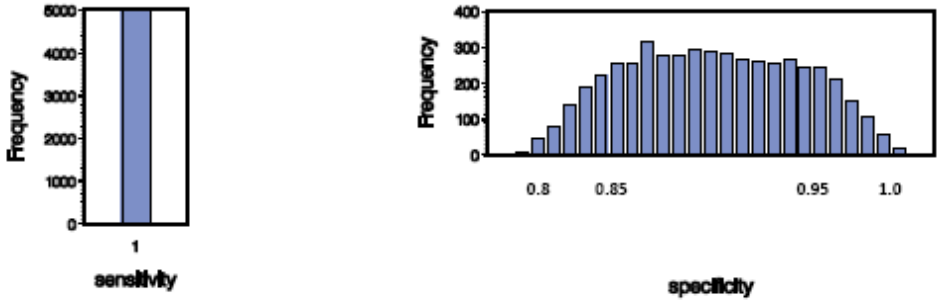


Figure 2 shows the output of sensitivity and specificity examples based on 5,000 iterations that are used to describe the patterns of nondifferential misclassification. For the random pattern, we specified a trapezoidal distribution (minimum of 80%, modes of 85 and 95%, and a maximum of 100%) for both sensitivity and specificity; for the upward pattern, we specified trapezoidal distribution (minimum of 80%, modes of 85 and 95%, and a maximum of 100%) for sensitivity and

a perfect specificity; while for the downward pattern, we specified a trapezoidal distribution (minimum of 80%, modes of 85 and 95%, and a maximum of 100%) for specificity and a perfect sensitivity.

Analysis and output

To conduct the probabilistic sensitivity analysis, we first randomly sampled a probability from each defined bias parameter distribution, followed by using the selected probability to correct for the outcome misclassification. After reclassification of each subject who was simulated to have been misclassified, a summary measure was estimated and saved to a reconstructed dataset. This reconstructed dataset represents a possible point estimate that could have occurred after correcting for misclassification, based on the probability distributions specified for sensitivity and specificity. The estimation of ordinal GOS was assessed using a proportional odds model and reported via a common odds ratio for the treatment versus the placebo group. No covariate was involved in the analysis.

For each study and each pattern of misclassification, the entire process described above was repeated 5,000 times to create a distribution of common odds ratios, which represented the corrected estimations (correcting for the misclassification only). The 95% simulation limits, which are the 2.5th and 97.5th percentiles of the corrected point estimates and the median estimate, were reported.

The study also took account of random error by calculating a standard error estimate for the log common odds ratio from the observed dataset, randomly choosing a standard normal deviate, and subtracting the product of this deviate and the standard error of the conventional point estimate. This process was also repeated 5,000 times for each reconstructed dataset, yielding a frequency distribution of odds ratios corrected for both systematic and random errors. Thus, three intervals were reported: the conventional 95% confidence limits (accounting for random error only), the

simulation limits that account for the misclassification only, and for both misclassification and random errors.

Results

Distribution of the 5-point GOS at the six month post injury

Table 1 shows the six month GOS outcome data that was used to explore the impact of ordinal outcome misclassification in TBI clinical trials. The outcome datasets were randomly sampled from eight RCTs and three observational studies contained in the IMPACT Database, representing the moderate to severe head injury population with either more restricted inclusion criteria (RCTs) or more generalized characteristics (observational studies). Each data set contains 800 patients with 400 in each arm. A 10% treatment effect was simulated for the treatment group to symbolize a trial effect.

Table 1. The Distribution of the Glasgow Outcome Scale at Six-month Post Injury in Traumatic Brain Injury Studies[1]

TBI Studies	Control group %				Treatment group %			
	Good recovery	Moderate disability	Severe disability	Vegetative status/Dead	Good recovery	Moderate disability	Severe disability	Vegetative status/Dead
Randomize controlled trials								
TINT	37.3	15.5	12.8	34.5	48.5	16.0	11.5	24.0
TIUS	42.3	19.0	12.5	26.3	51.0	18.5	12.3	18.3
SLIN	31.5	23.3	17.0	28.3	38.5	24.8	15.8	21.0
SAP	37.8	20.0	16.3	26.0	47.0	20.3	15.0	17.6
PEG	24.8	28.0	18.5	28.8	31.0	27.5	19.8	21.8
HITI	32.0	17.0	19.5	31.5	40.8	17.3	16.5	25.5
HITII	37.0	26.0	10.8	25.8	49.0	24.3	8.8	18.0
SKB	22.0	20.3	22.5	35.3	30.8	23.0	20.3	26.0
Observational studies								
TCDB	18.8	18.3	16.8	46.3	23.3	22.3	17.8	36.7
UK4	23.2	18.9	18.2	39.7	30.8	20.8	18.5	30.0
EBIC	30.0	17.3	16.3	36.5	37.3	20.8	15.3	26.8

1. Individual patient data randomly selected from the original TBI studies in the IMPACT database (Marmarou et al. 2007) with 10% simulated treatment effect, each dataset contains 800 patients with 400 in each arm.

In general, most studies had a 'U' shaped outcome distribution at the six month post injury, that is, large proportions of patients had outcomes of either good recovery or combined mortality and vegetative status, and relatively lower percentages of patients had outcomes of moderate or severe disabilities. The proportions of GR and VS/D among RCTs ranged from 22.0% (SKB) to 42.3% (TIUS) and from 25.5% (HITII) to 35.3% (SKB), respectively, while the proportions among the observational studies ranged from 18.8% (TCDB) to 30% (EBIC) and from 36.5% (EBIC) to 46.3% (TCDB), respectively. The proportions of MD and SD among the RCTs ranged from 15.5% (TINT) to 28% (PEG) and from 10.8% (HITII) to 19.5% (HITI), respectively, whereas the proportion among the observational studies ranged from 17.3% (EBIC) to 18.9% (UK4) and from 16.3% (EBIC) to 18.2% (UK4), accordingly.

As expected, the baseline GOS at the six month among the RCTs were better than the outcomes from the observational studies. The proportions of favorable outcome (GR and MD) were higher among the RCTs, compared with the proportions among the observational studies; whereas the mortalities were lower among the RCTs, as compared with the mortalities among the observational studies.

Probabilistic sensitivity analysis: correcting for nondifferential misclassification errors of ordinal GOS

Table 2 shows **t**he results of the ordinal GOS analyses by the conventional approach, as well as the probabilistic sensitivity analysis correcting for the nondifferential misclassification. The conventional analysis was performed on each observed dataset, from which a perfect outcome classification was assumed and 95% confidence intervals took account for random errors only. Among all studies, the common odds ratios of more favorable outcome, as compared between the

treatment and placebo group, ranged from 1.35 (PEG) (95% C.I., 1.06 - 1.71) to 1.62 (TINT) (95% C.I. 1.25 - 2.09).

Table 2. Results of the Six-month Ordinal GOS[1] Analysis
Comparison between the conventional approach and probabilistic sensitivity analysis correcting for nondifferential misclassification[2] with the specified sensitivity and specificity parameters

Analysis	Randomized controlled trials								Observational studies		
	TINT	TIUS	SLIN	SAP	PEG	HITI	HITII	SKB	TCDB	UK4	EBIC
Conventional approach	Estimated Common Odds Ratios (95% confidence intervals)										
	1.62 (1.25,2.09)	1.46 (1.13,1.89)	1.41 (1.09,1.81)	1.51 (1.17,1.95)	1.35 (1.06,1.71)	1.42 (1.11,1.82)	1.61 (1.23,2.08)	1.57 (1.22,2.00))	1.42 (1.12,1.85)	1.49 (1.16,1.92)	1.50 (1.18,1.91)
Probabilistic sensitivity analysis	Estimated Median Common Odds Ratios (2.5 percentile, 97.5 percentile of the simulation intervals)										
Correcting for misclassification with a random pattern[3]	1.67 (1.56,1.80)	1.51 (1.40,1.66)	1.44 (1.34,1.56)	1.55 (1.44,1.68)	1.39 (1.29,1.51)	1.44 (1.36,1.53)	1.71 (1.56,1.92)	1.60 (1.50,1.71)	1.43 (1.36,1.51)	1.51 (1.43,1.60)	1.50 (1.43,1.61)
Sensitivity and random error	1.67 (1.29,2.18)	1.51 (1.15,1.98)	1.45 (1.11,1.87)	1.55 (1.18,2.04)	1.39 (1.06,1.82)	1.43 (1.11,1.86)	1.71 (1.31,2.26)	1.60 (1.23,2.07)	1.43 (1.10,1.86)	1.51 (1.15,1.96)	1.50 (1.17,1.93)
Correcting for misclassification with an upward pattern[4]	1.66 (1.57,1.77)	1.49 (1.41,1.63)	1.42 (1.35,1.52)	1.53 (1.45,1.66)	1.36 (1.29,1.44)	1.44 (1.37,1.52)	1.66 (1.56,1.83)	1.59 (1.52,1.67)	1.43 (1.38,1.49)	1.50 (1.44,1.58)	1.50 (1.43,1.59)
Sensitivity and random error	1.66 (1.28,2.17)	1.50 (1.14,1.96)	1.43 (1.11,1.85)	1.53 (1.18,1.99)	1.36 (1.06,1.76)	1.44 (1.12,1.85)	1.67 (1.29,2.19)	1.59 (1.23,2.06)	1.43 (1.11,1.87)	1.50 (1.17,1.95)	1.50 (1.16,1.66)
Correcting for misclassification with a downward pattern[5]	1.43 (1.37,1.49)	1.28 (1.21,1.35)	1.12 (1.06,1.18)	1.24 (1.17,1.30)	1.01 (0.94,1.07)	1.13 (1.07,1.18)	1.45 (1.37,1.55)	1.13 (1.06,1.18)	1.13 (1.07,1.17)	1.17 (1.11,1.22)	1.22 (1.16,1.27)
Sensitivity and random error	1.43 (1.11,1.84)	1.28 (0.97,1.68)	1.12 (0.87,1.45)	1.24 (0.96,1.61)	1.02 (0.78,1.31)	1.13 (0.88,1.45)	1.46 (1.11,1.89)	1.13 (0.87,1.48)	1.12 (0.87,1.45)	1.16 (0.90,1.51)	1.21 (0.95,1.57)

1. The ordinal GOS are categorized as: D/VS, SD, MD and GR.
2. Nondifferential misclassification is assumed between the two adjacent outcome categories only except for the category of D/VS, for both treated and placebos.
3. Sensitivity and specificity are drawn from a trapezoidal distribution, with a minimum of 80% modes of 85% and 95% and a maximum of 100% for each.
4. Sensitivity is drawn from a trapezoidal distribution, with a minimum of 80% modes of 85% and 95% and a maximum of 100%, but specificity is defined as 100%.
5. Sensitivity is defined as 100%, but specificity is drawn from a trapezoidal distribution, with a minimum of 80% modes of 85% and 95% and a maximum of 100%.

Misclassification with random and upward patterns

The probabilistic sensitivity analysis correcting for the misclassification with a random pattern was demonstrated by a trapezoidal distribution specified for sensitivity and specificity, with a minimum 80%, modes of 85% and 95%, and a maximum of 100% for each. The 95% simulation limits, which are the 2.5th and 97.5th percentiles of the back calculated point estimates, and the corresponding median estimates moved upward slightly as compared to the results of the conventional approach for all studies. The actual simulation intervals (accounting for misclassification only) ranged from 1.29 - 1.51 (PEG) to 1.56 - 1.80 (TINT) and the corresponding median estimate ranged from 1.39 to 1.67. The overall 95% simulation limits (accounting for both

misclassification and random errors) ranged from 1.06 - 1.82 (PEG) to 1.29 – 2.18 (TINT) accordingly.

Given the specified sensitivity (minimum of 80%, modes of 85 and 95%, and a maximum of 100%) and specificity (100%) parameters, the analysis results correcting for the misclassification with an upward pattern were similar to the results correcting for the misclassification with a random pattern. The 95% simulation limits ranged from 1.29 - 1.44 (PEG) to 1.57 - 1.77 (TINT) and the corresponding median estimate ranged from 1.36 to 1.66. The overall 95% simulation limits ranged from 1.06 - 1.76 (PEG) to 1.28 – 2.17 (TINT) accordingly. If a random or upward pattern of nondifferential misclassification existed within the error ranges specified, the ordinal GOS from the observed datasets would have been underestimated by a small degree.

Misclassification with a downward pattern

In contrast, given the specification of the sensitivity (100%) and specificity (minimum of 80%, modes of 85 and 95%, and a maximum of 100%) parameters, the analysis results correcting for the misclassification with a downward pattern were quite different from the results correcting for the random and upward misclassification. The 95% simulation limits and the corresponding median estimate moved substantially downward for all studies, so did the overall 95% simulation limits for all studies. For the downward pattern, the 95% simulation intervals ranged from 0.94 – 1.07 (PEG) to 1.37 - 1.49 (TINT), and the corresponding median estimates ranged from 1.01 to 1.43. The overall simulation limits ranged from 0.78 – 1.31 (PEG) to 1.11 – 1.84 (TINT) accordingly. If a downward pattern of nondifferential misclassification existed within the assumed error ranges, the ordinal outcome from the observed datasets would have been inflated.

Discussion

We explored the impact of nondifferential misclassification on the 5-point ordinal scales among TBI studies using a probabilistic sensitivity analysis. The analysis involved reconstructing the data that would have been observed had the misclassified variable been correctly classified, given the sensitivity and specificity of classification. We have demonstrated that nondifferential misclassification could produce uncertainties for the 5-point ordinal GOS analysis in TBI trials. For instance, our simulation results showed that a) given a specification of a minimum of 80%, modes of 85% and 95% and a maximum of 100% (random pattern) for both sensitivity and specificity or b) given the same trapezoidal distributed sensitivity but a perfect specificity (upward pattern), the misclassification would have caused an ordinal GOS underestimated in the observed data drawn from the IMPACT database (Marmarou et al., 2007). In another scenario, given the same trapezoidal distributed specificity but a perfect sensitivity (downward pattern), the misclassification would have resulted in an inflated GOS estimation.

It is highly possible that the primary outcomes such as GOS and Extended GOS (GOSE) could have been misclassified to some extent in the TBI trials. Various studies have investigated misclassification and inter-observer variation of the TBI outcome measures and in general found that the variation does exist (Anderson et al., 1993; Brooks et al., 1986; Maas et al., 1983; Marmarou, 2001; Pettigrew et al., 1998; Scheibel et al., 1998; Teasdale et al., 1998; Wilson et al., 1998; Wilson et al., 2002; Wilson et al., 2007) The reported overall disagreement in GOS assessments ranged from 8% (Wilson et al., 1998) to 30% (Brooks et al., 1986); whereas the disagreement in GOSE ratings ranged from 22% (Wilson et al., 1998) to 41% (Wilson et al., 2007) in practices. When the overall disagreement in GOS assessment (collapsed from GOSE) from the study was broken down to individual categories (Wilson et al., 2007), the disagreement in rating the

categories of SD, MD and GR, between an expert and the untrained investigators, was 29.5%, 53.3% and 35% accordingly.

Three patterns of nondifferential misclassification

Our current analyses suggest that the scenarios of misclassification investigated are not unrealistic to clinical practice. Marmarou (2001) conducted a study among 34 American Brain Injury Consortium members to ascertain the reliability of the GOS rating. The results showed that the rating for 20.6% of Moderate patients was shifted to the Good Recovery GOS category and 32.3% of severe patients were rated as moderately disabled. An upward shift of outcome assignment had been previously reported (Anderson et al., 1993) and is a likely result of the optimism of the patient's primary care providers who compare the improved outcome to the serious condition immediately after injury, rather than to the healthy pre-injury status. Conversely, a rigid application of the criteria from the structured interview or questionnaires by research workers tends to allocate patients to lower outcome categories (Teasdale et al., 1998; Wilson et al., 1998). Therefore, nondifferential misclassification may be found in either the upward or downward direction, based on different clinical scenarios.

Correlation between the nondifferential misclassification and the probabilities of GOS categories

In this study, it appears that the impact of nondifferential misclassification on the 5-point ordinal GOS is less significant, compared with the effect of misclassification on the previously reported binary GOS (Lu et al., 2008). This is likely due to the probabilities or the prevalence of GOS categories which were misclassified. The examples of the correlation between the nondifferential misclassification and GOS category probabilities are given in Table 3. We propose that three GOS category probability sets (i.e., equal probability, the "U" shaped distribution and single dominant category) reflect the true outcome distribution, whereas the GOS assessment is

done with errors. The classification errors were illustrated via a simple model, in which 20% of patients in category GR were classified as being in MD, 20% MD being in GR, 20 % of patients in category MD being in SD, and 20 % SD being in MD for both placebo and treatment groups. As a result, the true category probabilities given at the beginning of each case are transformed by misclassification into the observed probabilities in the rows of 'Random misclassification.'

Table 3. Nondifferential Misclassification on the Probabilities of the GOS Categories

Cases	Analysis	Placebo (n=400)				Treatment (n=400)				OR[3]
		GR	MD	SD	D/VS	GR	MD	SD	D/VS	
Equal probability	True outcome[1]	0.25	0.25	0.25	0.25	0.33	0.27	0.22	0.18	1.50
	Random misclassification[2]	0.25	0.25	0.25	0.25	0.32	0.27	0.23	0.18	1.44
"U" shaped distribution	True outcome	0.35	0.15	0.15	0.35	0.45	0.15	0.14	0.27	1.50
	Random misclassification	0.31	0.19	0.15	0.35	0.39	0.21	0.14	0.27	1.46
Single dominant category	True outcome	0.20	0.50	0.20	0.10	0.27	0.51	0.15	0.07	1.50
	Random misclassification	0.26	0.38	0.26	0.10	0.32	0.39	0.22	0.07	1.35

1. The probabilities of the GOS categories were assumed as true, with a 10% treatment effect added to the treatment group based on the proportional odds model assumption.
2. For the random misclassification, a 20% upward and downward rate between the categories of SD and MD and between the categories of MD and GR was applied for both treatment groups.
3. The common odds ratio was estimated via a proportional odds model.

The results from our examples confirmed that the effect of misclassification on the cases of equal probability and "U" shaped distribution is relatively small. However, given the same error rates and treatment effect, the random misclassification caused the true outcome to be substantially underestimated in a single dominant GOS scenario, and the true outcome difference between placebo and treatment group reduced from 10% (OR=1.5) to 7.4% (OR=1.35). The scenario is similar with the effect of misclassification on binary GOS data. Thus, the impact of misclassification will likely be less sensible in the equal probability and the "U" shaped ordinal GOS distributions as observed in 11 TBI studies presented in Table 2.

Advantages and limitations of the probabilistic sensitivity analysis

Taken together, the scenario and the simulation intervals extrapolated by this study are in accordance with the previous study results. We applied a trapezoidal distribution to describe the misclassification parameters and patterns. The distribution is specified by four points: the lower (80%) and upper bounds (100%) and the lower (85%) and upper (95%) modes, between which the probability density is flat and equal to these modes, representing the zone of indifference. Thus, unlike the simple sensitivity analyses, the results from this probabilistic sensitivity analysis provide a sense of central tendency of the corrected ordinal GOS estimate. The results also provide a measure of uncertainty in the corrected estimate, as portrayed by the simulation limits. The confidence limits provided also include both classification and random errors. More significantly, our simulation study was based on data from eight major Phase III trials in TBI and three observational TBI studies. As such, we believe that our findings may be applicable to a wide range of trials in TBI.

It should be pointed out that, similar to all simulation studies, the main limitation of this study was that the distribution of the assumed misclassification parameter may be arbitrary, which could lead to different distributions of the adjusted analysis. Furthermore, the informed sensitivity analysis may be limited by the absence of any sense of weight to yield various results, such as the rate of misclassification between GR and MD or between MD and SD. In practice, the rate of misclassification may well be different between GR (good recovery) and MD (moderate disability) versus between MD and SD (severe disability).

Conclusions

In conclusion, the probabilistic sensitivity analysis in this study suggests that given the classification error ranges, the effect of nondifferential misclassification on the 5-point ordinal GOS

is likely to be small, compared with the impact on the binary GOS situation. The findings were consistent across eight major Phase III IBT trials and three observational studies. The results support the notion that the ordinal GOS analysis may not only gain the efficiency from the nature of the ordinal outcome, but also from the relative smaller impact of the potential misclassification, as compared with the conventional binary GOS analysis. Nevertheless, the outcome assessment following TBI is a complex problem. The assessment quality could be influenced by many factors. All possible aspects must be considered to ensure the consistency and reliability of the assessment and optimize the success of the trial.

Financial Disclosure: None

Acknowledgement: Grant Support was provided by NS-042691 and NS019235-21

References

Anderson, S.I., Housley, A.M., Jones, P.A., Slattery, J., and Miller, J.D. (1993). Glasgow Outcome Scale: an inter-rater reliability study. Brain injury : [BI]. 7, 309-317.

Bath, P.M., Geeganage, C., Gray, L.J., Collier, T., and Pocock, S. (2008). Use of ordinal outcomes in vascular prevention trials: comparison with binary outcomes in published trials. Stroke; a journal of cerebral circulation. 39, 2817-2823.

Brooks, D.N., Hosie, J., Bond, M.R., Jennett, B., and Aughton, M. (1986). Cognitive sequelae of severe head injury in relation to the Glasgow Outcome Scale. Journal of neurology, neurosurgery, and psychiatry. 49, 549-553.

Fox, M.P., Lash, T.L., and Greenland, S. (2005). A method to automate probabilistic sensitivity analyses of misclassified binary variables. International journal of epidemiology. 34, 1370-1376.

Jennett, B., and Bond, M. (1975). Assessment of outcome after severe brain damage. Lancet. 1, 480-484.

Lash, T.L., and Fink, A.K. (2003). Semi-automated sensitivity analysis to assess systematic errors in observational data. Epidemiology (Cambridge, Mass.). 14, 451-458.

Lu, J., Murray, G.D., Steyerberg, E.W., Butcher, I., McHugh, G.S., Lingsma, H., Mushkudiani, N., Choi, S., Maas, A.I., and Marmarou, A. (2008). Effects of Glasgow Outcome Scale misclassification on traumatic brain injury clinical trials. Journal of neurotrauma. 25, 641-651.

Maas, A.I., Braakman, R., Schouten, H.J., Minderhoud, J.M., and van Zomeren, A.H. (1983). Agreement between physicians on assessment of outcome following severe head injury. Journal of neurosurgery. 58, 321-325.

Marmarou, A. (2001). Head Trauma: Basic, Preclinical, Clinical Direction. First edn. Willey: New York, pps. 15

Marmarou, A., Lu, J., Butcher, I., McHugh, G.S., Mushkudiani, N.A., Murray, G.D., Steyerberg, E.W., and Maas, A.I. (2007). IMPACT database of traumatic brain injury: design and description. Journal of neurotrauma. 24, 239-250.

McCullaph, P. (1980). Regression-models for ordinal data. J. R. Statist. Soc. Ser. B Methodological 42, 109–142.

McHugh, G.S., Butcher, I., Steyerberg, E.W., Marmarou, A., Lu, J., Lingsma, H.F., Weir, J., Maas, A.I., and Murray, G.D. (2010). A simulation study evaluating approaches to the analysis of ordinal outcome data in randomized controlled trials in traumatic brain injury: results from the IMPACT Project. Clinical trials (London, England). 7, 44-57.

Optimising Analysis of Stroke Trials (OAST) Collaboration, Bath, P.M., Gray, L.J., Collier, T., Pocock, S., and Carpenter, J. (2007). Can we improve the statistical analysis of stroke trials? Statistical reanalysis of functional outcomes in stroke trials. Stroke; a journal of cerebral circulation. 38, 1911-1915.

Pettigrew, L.E., Wilson, J.T., and Teasdale, G.M. (1998). Assessing disability after head injury: improved use of the Glasgow Outcome Scale. Journal of neurosurgery. 89, 939-943.

Scheibel, R.S., Levin, H.S., and Clifton, G.L. (1998). Completion rates and feasibility of outcome measures: experience in a multicenter clinical trial of systemic hypothermia for severe head injury. Journal of neurotrauma. 15, 685-692.

Teasdale, G.M., Pettigrew, L.E., Wilson, J.T., Murray, G., and Jennett, B. (1998). Analyzing outcome of treatment of severe head injury: a review and update on advancing the use of the Glasgow Outcome Scale. Journal of neurotrauma. 15, 587-597.

Wilson, J.T., Edwards, P., Fiddes, H., Stewart, E., and Teasdale, G.M. (2002). Reliability of postal questionnaires for the Glasgow Outcome Scale. Journal of neurotrauma. 19, 999-1005.

Wilson, J.T., Pettigrew, L.E., and Teasdale, G.M. (1998). Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. Journal of neurotrauma. 15, 573-585.

Wilson, J.T., Slieker, F.J., Legrand, V., Murray, G., Stocchetti, N., and Maas, A.I. (2007). Observer variation in the assessment of outcome in traumatic brain injury: experience from a multicenter, international randomized clinical trial. Neurosurgery. 61, 123-8; discussion 128-9.

A METHOD FOR REDUCING MISCLASSIFICATION IN THE EXTENDED GLASGOW OUTCOME SCORE

Introduction

The eight-point extended Glasgow Outcome Scale (GOSE) was introduced (Jennett et al., 1981) to increase sensitivity of the primary outcome assessment in traumatic brain injury (TBI) trials. However, its assessment appears to be more complex and susceptible to inter-rater variation, as has been suggested by several sets of authors (Brooks et al., 1986; Maas et al., 1983; Marmarou, 2001), compared to the original version, the five-point Glasgow Outcome Scale (GOS; Jennett and Bond, 1975).

Conventionally, eight-point GOSE outcome data are collected through a structured interview with the patient, alone or together with a caretaker (Wilson et al., 1998). The structured interview is designed to reduce inter-rater variation through standardizing the questions relative to assessment, and to assist raters in recording the explicit reasons for classification into each GOSE category. Despite the fact that using the structured interview questionnaires helps reach acceptable agreement in GOSE assessment between raters (Pettigrew et al., 1998; Teasdale et al., 1998; Wilson et al., 1998), significant variation remains among different raters. A recent study using the structured interviews indicated an agreement rate as low as 59% (weighted kappa [κ]=0.72; 95% confidence interval [CI] 0.62, 0.75) for GOSE assessment by untrained investigators (Wilson et al., 2007).

Inter-rater variation in primary outcome rating is a serious concern that may have contributed to the lack of positive results in some TBI trials (Maas et al., 1999; Marmarou, 2001; Narayan et al., 2002). A study by Choi and colleagues (Choi et al., 2002) indicated that the effect of misclassification on GOS may not only decrease the desired power of a trial, but also the size of true benefit. Thus observer variation or outcome misclassification may obscure therapeutic effects

by introducing errors into the true study efficacy. We (Lu et al., 2008) recently reported that a 20% random misclassification on a dichotomous GOS outcome could reduce the treatment effect from the expected 10% to 6.8%, while maintaining the statistical power as a fixed factor.

The consistency and reliability of the outcome assessment could be influenced by many factors. Thus a collective effort by all possible means should be made to ensure the quality of the assessment. Here we introduce an alternate GOSE rating system as an aid in determining GOSE scores with the objective of reducing inter-rater variation in the primary outcome assessment in TBI trials.

The method used in this study is based on the concept that the GOSE is an extension of the GOS and as such, effort is focused on obtaining a reliable GOS score, and then limiting the questions asked in order to obtain a reliable GOSE score. More importantly, the method requires the investigator to record pre- and post-injury narratives to establish firm baselines, and source documentation that provides quality assurance through central monitoring to determine a reliable outcome and reduce clerical errors.

Materials and Methods

Study participation centers and design

Forty-five trauma centers in the United States were invited to participate in this study. These centers are members of the active American Brain Injury Consortium (ABIC) currently selected to participate in a Phase III TBI trial. The selection was based on the centers' past experience in TBI trials, the existing data regarding the annual volume of TBI patient enrollments, and the level of correspondence with ABIC. The selected centers were randomly divided into three study groups of equal size as balanced by the center's past experiences in TBI trials. These three groups were

assigned to use different methods to assess patient 6-month GOSE outcome as described in Figure 1.

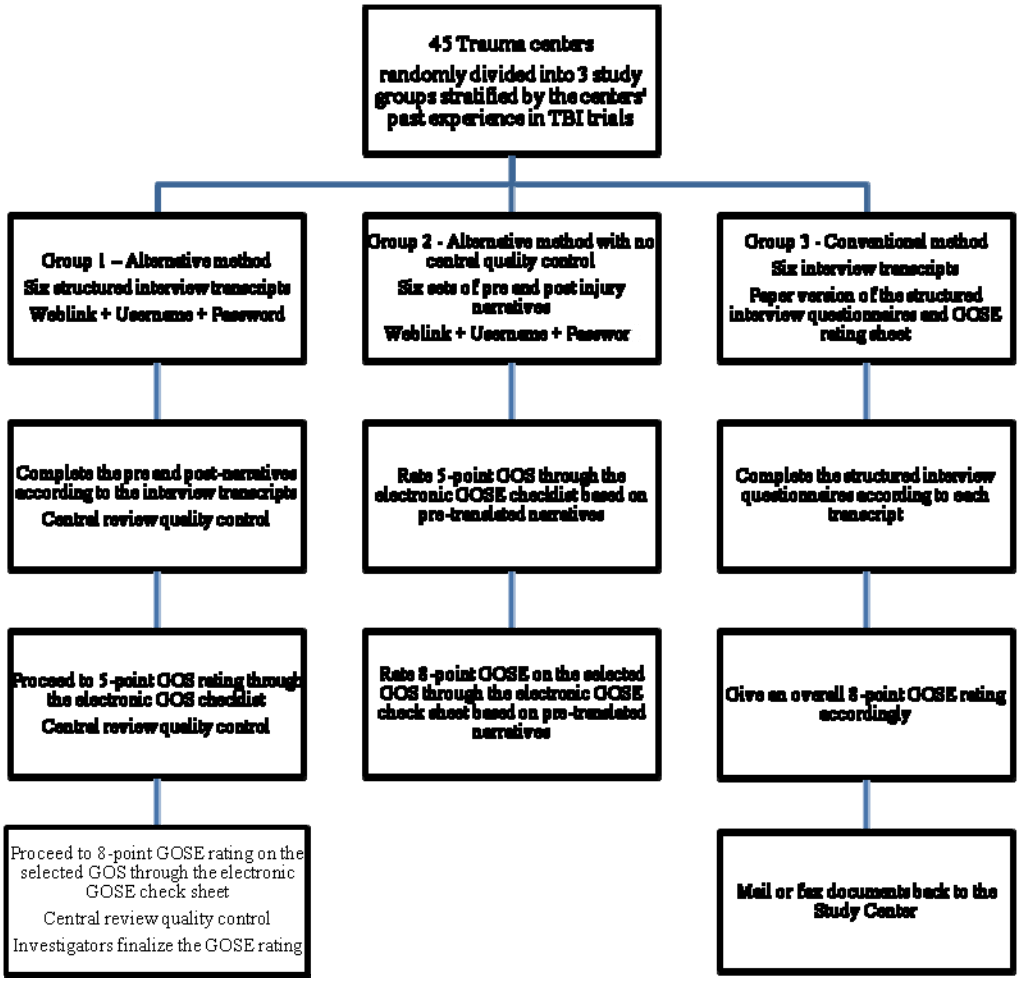


Figure 1.Forty-five trauma centers were randomly divided into three study groups balanced by each center's past experience in TBI trials. Group 1 used the alternative GOS/GOSE rating system coupled with central quality control, in which the raters were required to complete six sets of pre- and post-injury narratives according to six sample transcripts prior to outcome assessment. Group 2 used the alternative system with no central quality control, in which the raters used six sets of pre-specified narratives to rate the outcome. These narratives contained information, strictly transferred from the original interview transcripts by an expert, which allowed the validation of GOS/GOSE assessment without errors introduced by incorrect narratives. Group 3 used conventional structured interviews in which the raters were required to fill out the structured GOSE interview questionnaires based on the same six transcripts, and to provide an overall GOSE rating of the case (GOS, Glasgow Outcome Scale; GOSE, Extended Glasgow Outcome Scale; TBI, traumatic brain injury).

Group 1 used the alternative GOS/GOSE rating system coupled with central quality control, in which the raters were required to complete six sets of pre- and post-injury narratives according to six sample transcripts prior to the outcome assessment. Group 2 used the alternative system with no central quality control, in which the raters used six sets of pre-specified narratives to rate the outcome. These narratives contained information, strictly transferred from the original interview transcripts by an expert, which allowed the validation of GOS/GOSE assessment without errors introduced by incorrect narratives. Group 3 used conventional structured interviews in which the raters were required to fill out the structured GOSE interview questionnaires based on the same six transcripts, and to provide an overall GOSE score for the case. For each study group, the raters were given brief written instructions as to how to use the alternative system or conventional method to complete the outcome assessment. No additional training was given to the investigators. For study Group 1, the raters were informed that a central reviewer would monitor the rating process.

The alternative method was a web-based GOSE rating system, which required recording the structured pre- and post-injury narratives initially to establish firm baselines and source documentation. Based on the narratives, the system first captured the score on the five-point GOS according to six structured yes/no questionnaires. After the GOS category was defined, the system presented the raters with the criteria (Table 1) for the upper or lower strata of a particular GOS category in order to arrive at the GOSE. As such, only the questions relevant to the patient's GOS category were presented. For example, if the GOS was rated as moderate disability, the electronic system would route the rater to a screen where only questions regarding the upper and lower strata of moderate disability were presented. A set of the pre and post injury narratives and GOS and GOSE checklists is available as an online only supplement.

TABLE 1. GLASGOW OUTCOME SCALE (GOS) AND EXTENDED GLASGOW OUTCOME SCALE (GOSE)

Five-point GOS			*Eight-point GOSE*	
Category	*Key definition*[a]	*Key criteria*[b]	*Category*	*Key criteria*[b]
Good recovery (GR)	A patient is capable of resuming normal occupational and social activities with or without minor physical or mental deficits	1. Returns to work at the same level of performance as pre-injury ***and*** 2. Resumes at least more than half of the pre-injury level of social and leisure activities	Upper (GR+)	Returns to normal life with no current problems related to the head injury that affect daily life
			Lower (GR−)	1. Returns to pre-injury normal life, but has minor problems that affect daily life ***and/or*** 2. Resumes more than half the pre-injury level of social and leisure activities ***and/or*** 3. Disruption is infrequent (less than weekly)
Moderate disability (MD)	A patient is fully independent but disabled	1. Work capacity is reduced or unable to work ***and/or*** 2. Resumes less than half the pre-injury level of social and leisure activities	Upper (MD+)	1. Work capacity is reduced ***and/or*** 2. Resumes less than half the pre-injury level of social and leisure activities ***and/or*** 3. Disruption is frequent (once a week or more) but tolerable
			Lower (MD−)	1. Only able to work in sheltered workshop or unable to work ***and/or*** 2. Rarely or unable to participate in social/and leisure activities ***and/or*** 3. Disruption is constant (daily) and intolerable
Severe disability (SD)	A patient is conscious but needs the assistance of another person for some activities of daily living every day	1. Requires the help of someone to be around at home with activities of daily living ***and/or*** 2. Unable to travel or go shopping without assistance	Upper (SD+)	Can be left alone at least 8 h during the day, but unable to travel and/or go shopping without assistance
			Lower (SD−)	Requires frequent help of someone to be around at home most of the time every day
Vegetative status (VS)	Patient shows no evidence of meaningful responsiveness		Vegetative status (VS)	
Death (D)			Death (D)	

[a]Jennett et al., 1981.
[b]Wilson et al., 1998.

Moreover, a quality control system was built into the rating process that provided quality assurance through the use of a central reviewer. For instance, after the raters in Group 1 completed the pre- and post-injury narratives for each patient case, using information from the sample transcripts, a central reviewer would check whether the transferred narratives reflected accurate and sufficient information for assessing the outcome, compared with the original transcripts. The focus of the central review was to determine if there was sufficient information in all categories of the GOS/GOSE to arrive at an accurate assessment. Feedback from the central reviewer allowed the raters to re-check the narrative information if it was incomplete, or to proceed to the next step. The same quality control was performed after the raters completed each assessment of the five-point GOS and the eight-point GOSE, according to the raters' narratives. The investigators made the final decision based on the overall rating and the comments from the central review. Care was taken not to lead the investigators to a specific rating, but only to ensure that the information in the narrative was sufficient based on classic guidelines for GOS/GOSE assessment. In this way, the narratives served as a verifiable source document for the GOS and GOSE assessments.

Study material and outcome

Six transcripts of structured outcome interviews with patients with head injury or their relatives were used in order to assess the GOSE outcome. These transcripts contained real patient data originating from previous studies, and were also used in the dexanabinol study (Wilson et al., 2007) to assess baseline agreement between raters. The cases selected were not intended to be specifically representative of ''easy'' or ''difficult'' cases, but they covered the range of GOSE outcomes, from lower severe disability to lower good recovery, as assigned by an expert according to the criteria for the GOSE categories. The transcripts were distributed electronically to the study participating centers in two formats. For study Groups 1 and 3, the centers received the original

interview transcripts; for Group 2, the centers received six sets of pre-specified pre- and post-injury narratives that were transferred from the transcripts as described previously. No additional information regarding the outcome and the severity of injury were provided for these cases.

Statistical analysis

We analyzed the quality and inter-rater variation in GOS/GOSE assessment by the alternative GOSE data collection system (Groups 1 and 2). The results for outcome assessment were then compared against the results obtained using the conventional structured GOSE interviews (Group 3).

To identify whether central quality control played an important role in reducing inter-rater variation in the assessment of GOSE for the alternative method, we first applied the descriptive analyses and listed the discrepancies found in each step of the central quality check-ups for Group 1, including the steps of transferring patient responses from the original transcripts to pre-injury and post-injury narratives, and the assessments of the five-point GOS and the eight-point GOSE. We then compared the agreement rate in outcome ratings between the expert and the raters among all study groups. Further, to examine whether the two-stage GOSE assessment (i.e., assessing the five-point GOS first, then the eight-point GOSE) by the alternative system was more effective in reducing inter-rater variation, we compared the ratings for both five-point GOS and eight-point GOSE for all three groups through cross-tabulations.

The inter-rater agreement was assessed using weighted kappa (κ) statistics (Cohen, 1968). The weighted κ was developed to give more emphasis to the degree of disagreement. The conventional ‘‘weight’’ used for assessing disagreement in ordered categorical data was a quadratic weight. In general, the strength of agreement could be described by κ statistics as poor (<0.2), fair (>2 to _0.4), moderate (>0.4 to _0.6), good (>0.6 to _0.8), and very good (>0.8 to _1; Landis and

Koch, 1977). The weighted κ and its 95% confidence interval (CI), as well as the raw agreement rate were reported.

Results

Characteristics of the study centers

A total of 45 trauma centers were invited to participate and 32 centers volunteered to complete the study. The overall participation rate was 71%, and the participation rates for Groups 1, 2, and 3 were 67%, 73%, and 73%, respectively. The characteristics of the study participating centers and the raters' past experience in TBI trials and their current occupation status are described in Table 2.

TABLE 2. THE CHARACTERISTICS OF THE STUDY CENTERS BY GROUP

Characteristic	*Alternative system*	*Alternative system without central monitoring*	*Conventional structured interview*
Participation rate	67% (10/15)	73% (11/15)	73% (11/15)
Rater's past experience in TBI trials (*n*)	7 (7/10)	6 (6/11)	6 (6/11)
Rater's occupation status (*n*)			
Physician	3	2	3
Neuropsychologist		2	
Nurse	6	6	7
Other	1	1	1

The alternative GOSE rating system: Observation from central review

The analysis regarding the rule of central quality control for the alternative GOSE rating system was conducted for study Group 1. The raters completed three processes sequentially through the electronic rating system: pre/post narratives, GOS rating and GOSE rating. Ten raters each completed the three processes, including the transfer of information to the narratives, and rating of GOS followed by rating of GOSE for six cases. Out of 60 sample cases and 180 rating processes, the central reviewer identified 28 (28 out of 180) discrepancies, including 13 (13 out of 60)

discrepancies in the process of writing the post narratives, six (6 out of 60) in the five-point GOS assessment, and nine (9 out of 60) in the eight-point GOSE assessment. The investigators made the final decision on the overall rating, and the comments from central review resulted in rectifying 26 of the 28 discrepancies.

Major reasons for the discrepancies identified by the central reviewer for Group 1 are summarized in Table 3. Out of 13 discrepancies that occurred in the process of writing the post narratives, nine of those were because the raters did not respond to the specific questions that were required by the post narratives, while four cases were attributable to the raters' misinterpretation of the original information from the transcripts. For the six and nine discrepancies that were identified for the GOS and GOSE ratings, respectively, almost all discrepancies occurred because of incorrect outcome ratings based on the narratives. Namely, the narratives were correct, but the outcome rating was not in agreement with the narratives.

TABLE 3. DISCREPANCIES IDENTIFIED BY THE CENTRAL REVIEWER DURING THE OUTCOME RATING PROCESS FOR GROUP 1

	Overall discrepancies n = *180*	*Pre/post narrative set* n = *60*	*Five-point GOS* n = *60*	*Eight-point GOSE* n = *60*
Number (%) of discrepancies	28 (16)	13 (22)	6 (10)	9 (15)
Reasons for discrepancies				
Incorrect transfer of information from the transcripts (e.g., patient able to work, narrative says no)		4 cases		
The key criteria for GOSE assessment was missing in the narratives (e.g., more/less than half social activity)		9 cases		
Incorrect GOS/GOSE ratings based on the Narrative information			5 cases	9 cases
Other			1 case	
Number (%) of discrepancies corrected	26 (26/28)	12 (12/13)	6 (5/6)	9 (9/9)

GOS, Glasgow Outcome Scale; GOSE, extended Glasgow Outcome Scale.

Observer variation in assessment of the eight-point GOSE

The evaluation of consistency in eight-point GOSE assessment was conducted for all study groups as shown in Table 4a. For study Group 1, which was assigned to use the alternative GOSE rating system, the overall agreement in GOSE assessment between a central reviewer and the raters

was 97% (weighted κ=0.97; 95% CI 0.91, 1.00). This agreement rate was based on both investigators' overall rating and the central reviewer's comments. On two occasions the investigator disagreed with the comments from the central reviewer.

TABLE 4A. COMPARISON BETWEEN THE ALTERNATIVE EIGHT-POINT GOSE DATA COLLECTION METHOD AND THE CONVENTIONAL STRUCTURED INTERVIEWS: AGREEMENT BETWEEN A CENTRAL REVIEWER AND THE INVESTIGATORS ON RATING SIX SAMPLE CASE TRANSCRIPTS

			Investigator rating							
GOSE collection method	*Transcript*	*Expert*	*VS*	*SD−*	*SD+*	*MD−*	*MD+*	*GR−*	*GR+*	*Agreement*
Alternative system ($n=60$)	A	SD−		10						100%
	B	SD+			10					100%
	C	MD−				10				100%
	D	MD+					10			100%
	E	MD+			1		9			90%
	F	GR−					1	9		90%
Overall agreement 97% (weighted $\kappa=0.97$ and 95% confidence interval 0.91, 1.00)										
Alternative system without central monitoring ($n=66$)	A	SD−		10	1					92%
	B	SD+			11					100%
	C	MD−			1	5	5			45%
	D	MD+			2		9			82%
	E	MD+			2		9			82%
	F	GR−					4	6	1	55%
Overall agreement 76% (weighted $\kappa=0.79$ and 95% confidence interval 0.69, 0.89)										
Conventional structured interview ($n=66$)	A	SD−		6	5					55%
	B	SD+			9	2				82%
	C	MD−			1	6	2	2		55%
	D	MD+			1		7	3		64%
	E	MD+			1	4	6			55%
	F	GR−					2	7	2	64%

Overall agreement 63% (weighted $\kappa=0.70$ and 95% confidence interval 0.60, 0.81).
GR, good recovery; MD, moderate disability; SD, severe disability; VS, vegetative status; GOSE, extended Glasgow Outcome Scale.

Group 2 utilized the alternative rating system as well, but with no central quality control. The overall agreement rate in GOSE assessments between an expert and untrained raters was 76% (weighted κ=0.79; 95% CI 0.69, 0.89). In general, the raters did well in assessing the categories of lower and upper severe disabilities, for which the agreement rate between the central reviewer and raters reached 92% and 100%, respectively. However, the raters seemed to have more problems when assessing the sample cases of moderate disability and good recovery. The agreement rates in assessing the lower moderate GOSE equaled 45%, upper moderate GOSE 82%, and lower good GOSE 55%.

For the six lower moderate disability cases that an expert and the raters disagreed upon, five were due to the judgment of the patient's current occupational status and/or the degree of the social

and leisure activities resumed. In four disputed upper moderate disability cases, three were related to the inquiry as to whether the patients' current ability to drive or use public transportation was due to head injury or for some other reason. Finally, for five lower good recovery cases, four were not agreed upon as to whether the patient was able to return to their prior injury social and leisure activities by at least 50%.

Compared with the groups using the alternative rating system, the overall agreement between an expert and raters in Group 3 was lower. The overall agreement for Group 3 only reached 63% (weighted κ=0.70; 95% CI 0.60, 0.81). The agreement rates between an expert and the raters in assessing the categories were as follows: lower and upper severe disabilities (55% and 82%), lower and upper moderate disabilities (55% and 60%), and lower good recovery (64%). Moreover, the observed assessment disparity among the outcome categories was wider, especially in the assessment of moderate disabilities.

For the severe disability cases, except for one case of misunderstanding, six mistakes were due to algorithm issues. For the moderate disability categories, the majority of errors were in the area of social and leisure activities and/or current occupational status, for which the raters were required to exercise their own judgment in assessing if the social and leisure activities were more or less than 50%. Finally, the errors in rating the good recovery category were also seen mostly in the area of social and leisure activities.

Observer variation in assessment of the five-point GOS

The observer variation in the assessment of the five-point GOS is summarized in Table 4b. The performance on the five-point GOS rating scale was generally better among all study groups compared to the eight-point GOSE assessment. For Groups 1 and 2, that used the alternative approach to rate the outcome, the overall agreement between an expert and the raters were 97%

(weighted κ=0.95; 95% CI 0.89, 1.00), and 83% (weighted κ=0.81; 95% CI 0.69, 0.92), respectively. For Group 3, that used the conventional method, the overall agreement reached 83% (weighted κ=0.76; 95% CI 0.63, 0.89).

TABLE 4B. COMPARISON BETWEEN THE ALTERNATIVE FIVE-POINT GOS DATA COLLECTION METHOD AND THE CONVENTIONAL STRUCTURED INTERVIEWS: AGREEMENT BETWEEN A CENTRAL REVIEWER AND INVESTIGATORS ON RATING OF SIX SAMPLE CASE TRANSCRIPTS

		Investigator rating				
	Expert	*VS*	*SD*	*MD*	*GR*	*Agreement*
Alternative method ($n=60$)	SD		20			100%
	MD		1	29		97%
	GR			1	9	90%
Overall agreement 97% (weighted $\kappa=0.95$ and 95% confidence interval 0.89, 1.00)						
Alternative method without central monitoring ($n=66$)	SD		22			100%
	MD		5	28		85%
	GR			4	7	64%
Overall agreement 83% (weighted $\kappa=0.81$ and 95% confidence interval 0.69, 0.92)						
Conventional structured interview ($n=66$)	SD		20	2		91%
	MD		3	25	5	76%
	GR			2	9	82%

Overall agreement 83% (weighted $\kappa=0.76$ and 95% confidence interval 0.63, 0.89).
GR, good recovery; MD, moderate disability; SD, severe disability; VS, vegetative status; GOS, Glasgow Outcome Scale.

In accordance with the assessment of the eight-point GOSE, the raters did well on rating severe disability. The agreement rate between an expert and the raters for Groups 1 and 2 reached 100% and 100%, respectively, and the rate for Group 3 was 91%. However, the raters were less in agreement with the expert in the assessment of better GOS outcome categories. For Groups 1, 2, and 3, the agreement rates were 97%, 85%, and 76%, in the assessment of moderate disabilities, and 90%, 64%, and 82% in the assessment of good recovery, respectively.

Discussion

In this study, we used an alternative GOSE rating system to aid the assignment of outcome scores with the objective of reducing the inter-rater variation in the primary outcome assessment in TBI trials. The proposed system is an extension of the existing ABIC five-point GOS checklist, which was developed for the purpose of reducing inter-rater variation in GOS assessment in TBI trials (Wilson et al., 1998). The GOS checklist has been shown to decrease inter-observer variability

in a pilot trial (Marmarou, 2001), and was used in two TBI clinical trials (Marmarou et al., 1999, 2005). The current system adds additional criteria, while maintaining the five-point GOS rating criteria, to assess the use of the eight-point GOSE system, as directed by the guidelines (Wilson et al., 1998). In addition, the alternative system takes advantage of electronic data capture to (1) integrate a quality-control system into the rating process, which provides improved quality assurance through use of a central reviewer, (2) utilize an algorithm to arrive at the GOS score, and (3) to only present the questions separating the upper and lower categories of a specific GOS rating to arrive at the GOSE score.

The results of this study indicate that inter-rater variations in the outcome assessment can be reduced through the improved outcome data collection system. For study Group 1, which utilized the complete alternative system in which a central quality-control system was built into the rating process, the overall agreement rate between an expert and the raters in the assessment of the five-point GOS (weighted κ=0.95; 95% CI 0.89, 1.00), and the eight-point GOSE (weighted κ=0.97; 95% CI 0.91, 1.00), reached 97%. These results are superior to those of previous studies (Brooks et al., 1986; Maas et al., 1983; Marmarou, 2001; Wilson et al., 1998, 2007), as shown in Table 5.

Table 5. Agreement and Kappa in Previously Reported GOS/GOSE Assessments

Reference	*Methods*	*Overall agreement rate*		*Weighted kappa statistics (95% CI)*	
		GOS	*GOSE*	*GOS*	*GOSE*
Maas et al.,1983	1. Patient interview and sample cases	Structured questionnaires	60%	0.77 (0.55, 0.99)[a]	0.48 (0.28, 0.68)[a]
	2. Agreement between physicians	Patient interview: 86%	62%	0.71 (0.54, 0.88)[a]	0.52 (0.34, 0.70)[a]
		Sample cases: 81%			
Brooks et al.,1986	1. Patient cases	70%	46%	–	–
	2. Agreement between two experienced raters				
Wilson et al., 1998	1. Patient interview	Structured interview: 92%	78%	0.89 (–)	0.85 (–)
	2. Agreement between investigators				
Marmarou, 2001	1. Sample cases	GOS checklist 82%		–	
	2. Agreement between experts and investigators				
Wilson et al., 2007	1. Sample cases	Structured interviews:	59%	–	0.72 (0.68, 0.75)
	2. Agreement between experts and untrained investigators				
Lu et al., 2008	1. Sample cases	Alternative system with central	97%	0.95 (0.89, 1.00)	0.97 (0.91, 1.00)
	2. Agreement between experts and untrained investigators	review: 97%	63%	0.76 (0.63, 0.89)	0.70 (0.60, 0.81)
		Structured interview: 83%			

[a]Unweighted kappa statistics.
GOS, Glasgow Outcome Scale; GOSE, extended Glasgow Outcome Scale; CI, confidence interval.

Furthermore, the use of the alternative system alone (Group 2), without central monitoring, also demonstrated strength in lessening the variation in the eight-point GOSE assessment among untrained raters, especially in the assessment of lower and upper severe disability categories (Table 4a). The overall agreement, weighted κ, and CI [GOS 83%, 0.81, and (0.69, 0.92), GOSE 76%, 0.79, and (0.69, 0.89)] in the outcome assessment were better than the results reported earlier, and consistent with more recent results (Marmarou, 2001; Wilson et al., 1998, 2007). The results from Group 3, that used the conventional structured interviews, are in close agreement with the baseline variability found in the dexanabinol trial. (Wilson et al., 2007).

Moreover, this study provided valuable insights into (1) potential causes of inter-rater variations during the outcome assessment process, and (2) the impact of an improved outcome rating system on constraining such variations in the course of assessment. The proposed system may help reduce the variation in the assessment of the eight-point GOSE through the following approaches.

Pre-injury narratives

The first step in this alterative system was to collect a pre-injury narrative within 2 weeks post-injury. This helped in determining the true impact of head injury on an individual's daily functioning, by taking into consideration the pre-injury status for each of the areas included in the GOS and GOSE assessment scales. Given the performance of Group 1 raters, the description and format of the pre-injury narrative appeared to be sufficient to serve as an important baseline reference. In comparison with sample cases, no disagreement with the narratives was shown between the central reviewer and raters. Thus it appears that these narratives in the module are user-friendly and self-exploratory for future use in TBI trials.

Three- and six-month post-injury narratives

Since this alternative system requires an investigator responsible for the outcome assessment to collect the data for the 3- and 6-month post-injury narratives after documenting patient function as described above for the pre-injury narratives, this indicates that each patient serves as their own control. As such, a reduction in GOS score will more clearly reflect the result of the patient's head injury. These narratives provide not only the necessary information for the outcome rating, but also critical source documentation for the purpose of quality control.

In this study, we found that most raters were able to properly record the post-injury information according to the descriptions of the post-injury narratives. Out of 13 post-injury narratives that the central reviewer had to query, nine cases were due to the raters' failure to respond to the questions required, and four were caused by the raters' misinterpretation of sample cases. At least two main reasons may explain these errors. For instance, the raters' prior knowledge and experience in utilizing a rating instrument may be an important factor. In the absence of knowledge of the basic concepts and understanding of the outcome and rating process, a rater does not possess the information necessary to assess the outcome, even when the information is available, as suggested by the results of several previous studies. In this regard, Clifton and colleagues (2001) demonstrated that higher-enrollment centers were superior in data completion, outcome assessment, and overall patient management, compared to lower-enrollment centers. Wilson and associates (2007) also showed that the central review identified a relatively large number of discrepancies (29–37%) during the early stages of a trial, but the number declined as the trial progressed, which coincided with more extensive investigator training and feedback from the central review.

Moreover, a loosely-structured question format leads to an open-ended answer and flexibility for the raters to provide their answers. As such, using a mixture of open-ended and pre-

categorized answers would be expected to improve data collection. For example, loosely-structured more open-ended answers allows raters to better document an individual's post-injury condition, while the fixed pre-categorized answers facilitates a standardized outcome rating among the patient population. This may be particularly meaningful with regard to the key concepts that differentiate the GOS and GOSE categories, as shown in Table 1.

It should be pointed out that in this study, the raters were able to obtain feedback from a central reviewer and had the opportunity to correct an error in narrative writing before the rating process. Thus, we recommend (1) acquiring the necessary knowledge about the outcome and its assessment instrument prior to a trial, (2) practicing on sample cases before the actual assessment is carried out, (3) collecting complete information in accordance with the requirements outlined in the narratives, and (4) checking the consistency of their own narratives.

A two-stage GOS and GOSE rating system

The alternative GOSE rating system requires the investigators to rate the less complicated five-point GOS first, followed by rating the eight-point GOSE category, by subdividing the selected GOS category into a ''lower'' or ''upper'' category. The rating on the five-point GOS is based on a checklist that contains the same category and layout of the source documentation (i.e., the pre- and post-injury narratives). But the eight-point GOSE rating is based on a compilation sheet in which the information is extracted from the pre-categorized answers of the post-injury narrative sheet. *Thus, this system provides a two-stage rating process, thereby minimizing potential observer variations across the five-point GOS outcome categories, and simplifies the assessment of the eight-point GOSE.*

To date, this system has been used to collect 6-month GOSE data for an observational study, including both U.S and European centers. In this validation study, we found that a two-stage outcome rating system was, in general, an improvement over the conventional approach in the eight-point GOSE assessment. The improvement was particularly noticeable in the assessment of severe disability. For instance, of 22 lower and upper severe disability sample cases, the expert had only one disagreement with the ratings obtained from Group 2, in which the two-stage system was used, whereas the expert had seven disagreements (7/22) from Group 3, in which the conventional method was applied. It seems that the more reliable rating outcome for Group 2 was directly related to the use of the two-stage system, which simplified the rating process and automated the rating algorithms. This is especially evident in light of the fact that neither group received feedback from the central quality control, and both used the same sample cases.

Study limitations

Although studies of inter-rater variations using central reviews or ratings from sample cases can reveal inconsistencies in outcome assessment, they are unlikely to capture every potential type of variation. In practice, the outcome assessment may be more complex, and the results may be further influenced by how the questions are asked and responses are solicited. Thus, the inter-observer agreements obtained in this study, based on sample transcripts, cannot be directly extrapolated to the clinical situation when assessing actual patients. Also, the results of this study were obtained from a relatively small group of investigators. Further study with larger groups of investigators in actual interview situations are needed to further confirm the results of this study. Moreover, the method of using case histories does not allow further information to be gathered over what has already been collected in the sample cases.

Nevertheless, since the sample cases used in this study were originally obtained through the structured interviews, and used in a large Phase III head injury trial GOSE inter-observer variation study (Wilson et al., 2007), it was reasonable to believe that the information obtained from these cases included sufficient information to assess patient outcome. Therefore, we believe that these sample cases are useful to validate whether (1) the criteria described in the narratives provide sufficient information for outcome assessment, and (2) the alternative GOSE rating system itself is better at reducing variations in GOSE assessment than conventional structured interviews.

Conclusion

The results of this study indicate that the alternative method for GOSE assessment has several advantages over current techniques. First, a narrative provides source documentation about the pre-injury status, and the status at 3 and 6 months post-injury, thus allowing for a more thorough central review. Second, a GOS-structured checklist provides an easy and practical method for GOS assessment. Third, an electronic system that directs the investigator to focus on an upper or lower classification of the GOSE criteria provides an easy and practical method for GOSE assessment. Taken together these elements, coupled with the central review, allow a more reliable GOSE rating system, thus reducing inter-rater variation and misclassification. The results of this study emphasize the importance of combining all efforts to reduce outcome misclassification, including the use of a reliable outcome rating system, collection of sufficient standardized information, proper rater training, and central quality control.

Author Disclosure Statement

No competing financial interests exist.

Acknowledgment

Support for this work was provided by National Institutes of Health grant NS 42691.

Participating Study Centers

Allegheny General Hospital, Christiana Care Health Services, Froedtert Memorial Lutheran Hospital=MCW, Harbor-UCLA Medical Center, Hennepin County Medical Center, John Peter Smith Hospital, Legacy Health System, Loyola University Medical Center, Louisiana State University Health Sciences Center-Shreveport, Maricopa Integrated Health System, Miami Valley Hospital, Mount Sinai Hospital Medical Center-Chicago, St. Louis University Hospital, Southern Illinois University School of Medicine, Springfield Neurological and Spine Institute, University of Medicine and Dentistry of New Jersey University Hospital, University of California-Davis Medical Center, University of Wisconsin Hospitals and Clinics, University of Cincinnati Medical Center, University of Iowa Hospitals and Clinics, University of Miami, University of Mississippi Medical Center, University of New Mexico, University of Oklahoma Health Sciences Center, University of Pennsylvania Medical Center,

University of Pittsburgh Medical Center, University of Tennessee Health Sciences Center at Memphis, University of Utah Health Sciences Center, University of Virginia Health System, Virginia Commonwealth University Medical Center, West Virginia University Hospitals, Wishard Memorial Hospital.

References

Brooks, D.N., Hosie, J., Bond, M.R., Jennett, B., and Aughton, M. (1986). Cognitive sequelae of severe head injury in relation to the Glasgow Outcome Scale. J. Neurol. Neurosurg. Psychiatry 49, 549–553.

Choi, S.C., Clifton, G.L., Marmarou, A., and Miller, E.R. (2002). Misclassification and treatment effect on primary outcome measures in clinical trials of severe neurotrauma. J. Neurotrauma 19, 17–22.

Clifton, G.L., Choi, S.C., Miller, E.R., Levin, H.S., Smith, K.R., Jr., Muizelaar, J.P., Wagner, F.C., Jr., Marion, D.W., and Luerssen, T.G. (2001). Intercenter variance in clinical trials of head trauma—experience of the National Acute Brain Injury Study: Hypothermia. J. Neurosurg. 95, 751–755.

Cohen, J. (1968). Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. Psychol. Bull. 70, 213–220.

Jennett, B., and Bond, M. (1975). Assessment of outcome after severe brain damage. Lancet 1, 480–484.

Jennett, B., Snoek, J., Bond, M.R., and Brooks, N. (1981). Disability after severe head injury: observations on the use of the Glasgow Outcome Scale. J. Neurol. Neurosurg. Psychiatry

44, 285–293.

Landis, J.R., and Koch, G.G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. Biometrics 33, 363–374.

Lu, J., Murray, G.D., Steyerberg, E.W., Butcher, I., McHugh, G.S., Lingsma, H., Mushkudiani, N., Choi, S., Maas, A.I., and Marmarou, A. (2008). Effects of Glasgow Outcome Scale misclassification on traumatic brain injury clinical trials. J. Neurotrauma 25, 641–651.

Maas, A.I., Braakman, R., Schouten, H.J., Minderhoud, J.M., and van Zomeren, A.H. (1983). Agreement between physicians on assessment of outcome following severe head injury. J. Neurosurg. 58, 321–325.

Maas, A.I., Murray, G., Henney, H., 3rd, Kassem, N., Legrand, V., Mangelus, M., Muizelaar, J.P., Stocchetti, N., Knoller, N., and Pharmos TBI investigators. (2006). Efficacy and safety of dexanabinol in severe traumatic brain injury: results of a phase III randomised, placebo-controlled, clinical trial. Lancet Neurol. 5, 38–45.

Maas, A.I., Steyerberg, E.W., Murray, G.D., Bullock, R., Baethmann, A., Marshall, L.F., and Teasdale, G.M. (1999). Why have recent trials of neuroprotective agents in head injury failed to show convincing efficacy? A pragmatic analysis and theoretical considerations. Neurosurgery 44, 1286–1298.

Marmarou, A., Guy, M., Murphey, L., Roy, F., Layani, L., Combal, J.P., Marquer, C., and American Brain Injury Consortium. (2005). A single dose, three-arm, placebo-controlled, phase I study of the bradykinin B2 receptor antagonist Anatibant (LF16-0687Ms) in patients with severe traumatic brain injury. J. Neurotrauma 22, 1444–1455.

Marmarou, A. (2001). Head Trauma: Basic, Preclinical, Clinical Direction, 1st ed. Wiley: New York, p. 15.

Marmarou, A., Nichols, J., Burgess, J., Newell, D., Troha, J., Burnham, D., and Pitts, L. (1999). Effects of the bradykinin antagonist Bradycor (deltibant, CP-1027) in severe traumatic brain injury: results of a multi-center, randomized, placebo-controlled trial. American Brain Injury Consortium Study Group. J. Neurotrauma 16, 431–444.

Narayan, R.K., Michel, M.E., Ansell, B., Baethmann, A., Biegon, A., Bracken, M.B., Bullock, M.R., Choi, S.C., Clifton, G.L., Contant, C.F., Coplin, W.M., Dietrich, W.D., Ghajar, J., Grady, S.M., Grossman, R.G., Hall, E.D., Heetderks, W., Hovda, D.A., Jallo, J., Katz, R.L., Knoller, N., Kochanek, P.M., Maas, A.I., Majde, J., Marion, D.W., Marmarou, A., Marshall, L.F., McIntosh, T.K., Miller, E., Mohberg, N., Muizelaar, J.P., Pitts, L.H., Quinn, P., Riesenfeld, G., Robertson, C.S., Strauss, K.I., Teasdale, G., Temkin, N., Tuma, R., Wade, C., Walker, M.D., Weinrich, M., Whyte, J., Wilberger, J., Young, A.B., and Yurkewicz, L. (2002). Clinical trials in head injury. J. Neurotrauma 19, 503–557.

Pettigrew, L.E., Wilson, J.T., and Teasdale, G.M. (1998). Assessing disability after head injury:

improved use of the Glasgow Outcome Scale. J. Neurosurg. 89, 939–943.

Teasdale, G.M., Pettigrew, L.E., Wilson, J.T., Murray, G., and Jennett, B. (1998). Analyzing outcome of treatment of severe head injury: a review and update on advancing the use of theGlasgow Outcome Scale. J. Neurotrauma 15, 587–597.

Wilson, J.T., Pettigrew, L.E., and Teasdale, G.M. (1998). Structured interviews for the Glasgow Outcome Scale and the extended Glasgow Outcome Scale: guidelines for their use. J. Neurotrauma 15, 573–585.

Wilson, J.T., Slieker, F.J., Legrand, V., Murray, G., Stocchetti, N., and Maas, A.I. (2007). Observer variation in the assessment of outcome in traumatic brain injury: experience from a multicenter, international randomized clinical trial. Neurosurgery 61, 123–128; discussion 128–129.

APPENDIX: Pre- and Post-Injury Narratives, GOS and GOSE Checklist

TBI STUDY	American Brain Injury Consortium Identification Trial # | Center # | Subject # | Subject Initials	PRE-INJURY NARRATIVE

PRE-INJURY NARRATIVE – BASELINE STATUS

Occupational Status:

Pre-injury employment, academic, volunteer or household responsibilities: describe their roles and status, number of hours worked each week or credit hours at school, level of performance etc. If the subject is currently unemployed, describe their history and any unusual circumstances.

Social Skills:

Describe where the patient lived before the injury, their family roles and responsibilities, level of participation in family decisions, dating/marital status, number of children and quality of each relationship, other relatives, friends and the quality of each relationship, hobbies and level of participation, leisure activities and frequency, such as dining out, movies or social gatherings or describe how the subject spent their free time.

Activities of Daily Living:

Describe the level of independence at home in routine activities of daily living such as feeding, toileting, grooming, (bathing, showering, brushing teeth) and basic hygiene, etc; the level of daily activities and household responsibilities such as preparing simple meals, managing money, and other household chores such as cleaning, laundry, cutting grass etc.

Mode of Transportation:

Describe the level of independence outside home with transportation, such as driving a car, using public transportation, taxi etc., or describe the pre-injury situation, example - if the subject has lost their license or used to ride the bus until they moved to the country. It is important to describe their "ability" to independently transport themselves in some circumstances.

General:

Describe any pre-injury physical deficits, mental deficits or common complaints.

The information documented above is based upon an interview with (check all that apply):

☐ Patient alone ☐ Patient plus relative/friend/caretaker ☐ Medical Personnel ☐ Relative/friend/caretaker alone

Date & Time of Interview:

DD / MM / YY HH : MM Interviewer Name/Signature Title/Profession

TBI Study	American Brain Injury Consortium Identification Trial # Center # Subject # Subject Initials	3/6-Month GOS/GOSE Narrative

3/6-MONTH GOS/ GOSE NARRATIVE

Compare subject's current status to their Pre-Injury/Baseline status.

Follow-up at: O 3-Month **or** O 6-Month

Occupational Status:

Current employment, academic, volunteer or household responsibilities: describe their roles and status, number of hours worked each week or credit hours at school, level of performance etc. If the subject is currently unemployed, describe their history and any unusual circumstances. _____

_____

_____

Summarize subject's work capacity:

- O Returned to normal level
- O Reduced
- O Only able to work in sheltered workshop or unable to work

Social Skills:

Describe where the patient currently lives, family roles and responsibilities, level of participation in family decisions, dating/marital status, number of children and quality of each relationship, other relatives, friends and the quality of each relationship, hobbies and level of participation, leisure activities and frequency, such as dining out, movies or social gatherings or describe how the subject spends their free time. _____

_____

_____

Has the subject resumed pre-injury level of social activities

- O At least half as often
- O Less than half as often
- O Unable to or rarely participate

Disruption or strain:

Describe if there are problems in family or friendship disruption due to psychological problems. _____

_____

_____

Summarize the subject's extent of disruption or strain:

- O Occasional (less than once per week)
- O Frequent (once per week or more but tolerable)
- O Constant (daily and intolerable)

TBI Study	American Brain Injury Consortium Identification Trial # Center # Subject # Subject Initials	3/6-Month GOS/GOSE Narrative

Activities of Daily Living:

Describe the level of independence at home in routine activities of daily living such as feeding, toileting, grooming, (bathing, showering, brushing teeth), preparing simple meals, remembering to take medication and reacting appropriately in case of an emergency, etc.

_____

_____

Are the family members able to leave the subject at home alone for at least 8 hours during the day, if necessary?

- O Yes
- O No

Mode of Transportation:

Describe the current level of independence outside home using transportation and being able to shop. Example – such as driving a car, using public transportation, calling a taxi, giving instructions to the driver, planning and managing money, and behaving appropriately in public.

_____

_____

_____

Is the subject able to drive or use public transportation, such as the bus or call a taxi, without assistance?

- O Yes
- O No

Is the subject able to shop without assistance?

- O Yes
- O No

General:

Describe any current physical deficits, mental deficits or common complaints. _____

_____

_____

Are there any problems (dizziness, headaches, sensitivity to noise or light, slowness, memory failure and concentration problems) that affect daily life after the head injury?

- O Yes
- O No

The information documented above is based upon an interview with (check all that apply):

- ☐ Patient alone
- ☐ Patient plus relative/friend/caretaker
- ☐ Medical Personnel
- ☐ Relative/friend/caretaker alone

Date & Time of Interview:

DD / MM / YY HH : MM Interviewer Name/Signature Title/Profession

TBI Study	American Brain Injury Consortium Identification Trial # Center # Subject # Subject Initials	3/6-Month GOS Checklist

3/6-MONTH GLASGOW OUTCOME SCALE CHECKLIST

Interview Date: _____ (dd/mm/yy)	GOS Assessed By:	Title/Profession:
Follow-up Month: O 3-Month **or** O 6-Month	GOS Confirmed By:	Title/Profession:

INSTRUCTIONS: Interview the subject, family members, friends and/or rehabilitation caregivers at three and six months after injury to obtain information relating to the following areas. Remember to always ask WHY the patient is or is not involved in certain pre-injury activities. Items should only be rated as "No" if subjects have not resumed their pre-injury activities due to physical or mental deficits that are *A RESULT OF THE BRAIN INJURY.* (For example, if a patient is safe to be at home alone, but is not left alone because parents/spouse are overprotective, they should still be rated as independent (Yes) in that area. A second example, if a patient is disabled because of a crush injury to the leg and that is the only reason they cannot return to climbing ladders at work, they should still be rated as (Yes), since the disability is not a "Brain Related Injury". Also note, if a subject was not able to perform certain tasks prior to the injury, do not expect them to perform those activities after the injury).

The information documented above is based upon an interview with (check all that apply):
☐ Patient alone ☐ Patient plus relative/friend/caretaker ☐ Medical Personnel ☐ Relative/friend/caretaker alone

Please answer all the questions	Yes	No
1. HAS THE SUBJECT RESUMED PRE-INJURY OCCUPATIONAL STATUS? If yes, check one box on the left, A, B, C or D, that is most applicable to the subject. ☐ A. Subject has resumed pre-injury **employment**: (e.g. executive or skilled laborer that has returned to the same level of responsibility at the same number of hours worked each week at the same performance level equivalent to that of their pre-injury status, with or without the use of compensatory mechanisms.) ☐ B. Subject could have resumed pre-injury occupational status, but currently has a reduction in performance level due to **outlying circumstances** that are not brain related; (e.g. previous work place out of business so they are unemployed and job hunting or a physical injury such as a leg amputation that prevents them from returning to work – these are not "Brain Related Injuries" and should not be rated as a No for these types of reasons.) ☐ C. Subject has resumed **academic** pursuits: (e.g. a student has resumed the same number of hours/classes at the same performance level equivalent to that of their pre-injury abilities - independently, without the need for special accommodations such as tutors, if they were not necessary prior to the injury.) ☐ D. Subject has resumed **household** responsibilities; (e.g. a subject that was previously a housewife, househusband, unemployed and/or volunteer that has returned to the same level of responsibility, same number of hours involved at the same performance level equivalent to that of pre-injury.)		
2. HAS THE SUBJECT RESUMED AT LEAST HALF AS OFTEN AS THE PRE-INJURY LEVEL OF SOCIAL ACTIVITIES? In areas such as family roles, responsibilities and relationships, friendships or established new ones and has resumed pre-injury hobbies, regular participation in leisure activities such as attending movies and dining out.		
3. HAS THE SUBJECT RESUMED PRE-INJURY LEVEL OF INDEPENDENCE IN ACTIVITIES OF DAILY LIVING? In areas such as feeding, toileting, grooming (bathing/showering, brushing teeth) and basic hygiene etc.		
4. IS THE SUBJECT ABLE TO BE LEFT ALONE AT HOME AND IS ABLE TO CARE FOR HER/HIM SELF AT THE SAME LEVEL OF PRE-INJURY? Specifically in areas such as the preparation of meals, necessary housework remembers to take medications and reacts appropriately in the case of an emergency, etc		
5. HAS THE SUBJECT RESUMED INDEPENDENCE IN TRANSPORTATION? Subject is able to drive or use public transportation, such as the bus or call a taxi, without assistance.		
6. DOES THE SUBJECT RESPOND TO VERBAL COMMUNICATION AND/OR OBEYS COMMANDS?		
RATING THE GOS: Review the Yes/No responses above and check (Y) only one GOS rating box. ☐ GOOD RECOVERY — If you answered YES to ALL questions ☐ MODERATE DISABILITY — If you answered NO to question 1 &/or 2, and YES to 3, 4, 5 & 6. ☐ SEVERE DISABILITY — If you answered NO to question 3, 4 or 5 and YES to 6. ☐ VEGETATIVE — If you answered NO to question 6		

TBI Study	American Brain Injury Consortium Identification Trial # Center # Subject # Subject Initials	3/6-Month **Extended** GOS Checklist

3/6-MONTH EXTENDED GLASGOW OUTCOME SCALE

COMPLETE THIS FORM ONLY AFTER ESTABLISHING GOS

Interview Date: ☐☐/☐☐/☐☐ (DD MM YY)	**GOSE Assessed By:**	**Title/Profession:**
	GOSE Confirmed By:	**Title/Profession:**
Source of Information: ☐ Patient alone ☐ Relative/friend/caretaker alone ☐ Patient plus relative/friend/caretaker ☐ Medical Personnel		**Follow-up Month:** O 3-Month **or** O 6-Month

	Upper Good	**Your Evaluation**
GOOD	*Returned to normal life* and *NO current problems* relating to the injury that affect daily life (dizziness, headache, and sensitivity to noise or light, slowness, memory failure, concentration problem)	☐
	Lower Good	
	Returned to normal life BUT current problems relating to the injury that affect daily life (dizziness, headache, sensitivity to noise or light, slowness, memory failure, concentration problem)	☐
	Social activities: Resumed at least half as often as pre-injury	
	Disruption or Strain: Occasional (less than once per week)	

	Upper Moderate	**Your Evaluation**
MODERATE	*Work capacity:* Reduced	☐
	Social activities: Resumed less than half as often as pre-injury	
	Disruption or Strain: Frequent (once a week or more but tolerable)	
	Lower Moderate	
	Work: Unable to work or only able to work in sheltered workshop	☐
	Social activities: Unable or rarely to participate	
	Disruption or Strain: Constant (daily and intolerable)	

	Upper Severe	**Your Evaluation**
SEVERE	*Care:* Does not require someone to be around at home most of the time (for at least 8-hours during the day, if necessary)	☐
	Travel: Unable to travel locally without assistance	
	Shopping: Unable to shop without assistance	
	Lower Severe	
	Care: Requires frequent help of someone to be around at home most of the time	☐

Other typical problems reported after head injury: headaches, dizziness, sensitivity to noise or light, slowness, memory failures and concentration problems.

If similar problems were present before the injury have these become markedly worse? ☐ Yes ☐ No

What is the most important factor in outcome?

- ☐ Effects of head injury
- ☐ Effects of illness or injury to another part of the body
- ☐ A mixture of these

Is or has the patient been in organized rehabilitation? ☐ Yes ☐ No

Vita

Juan Lu was born on April 29, 1958, in Suzhou, Jiangsu Province, China. She received her MD degree in 1983 from the School of Medicine of Suzhou University in China and completed her residency training in obstetrics and gynecology from The First Clinical College of Nanjing Medical University in 1986. After practicing for a few years she immigrated to the United States of America in 1990 and is now a US citizen. She acquired her master's degree in public health from the Virginia Commonwealth University in 2001. For the past ten years, she has worked at the American Brain Injury Consortium and Department of Neurosurgery of the Virginia Commonwealth University Health System, where she currently is a senior statistical manager and research project manager. While working full time, she finished her Ph.D. training in epidemiology with the Department of Epidemiology and Community Health of the Virginia Commonwealth University in 2010.